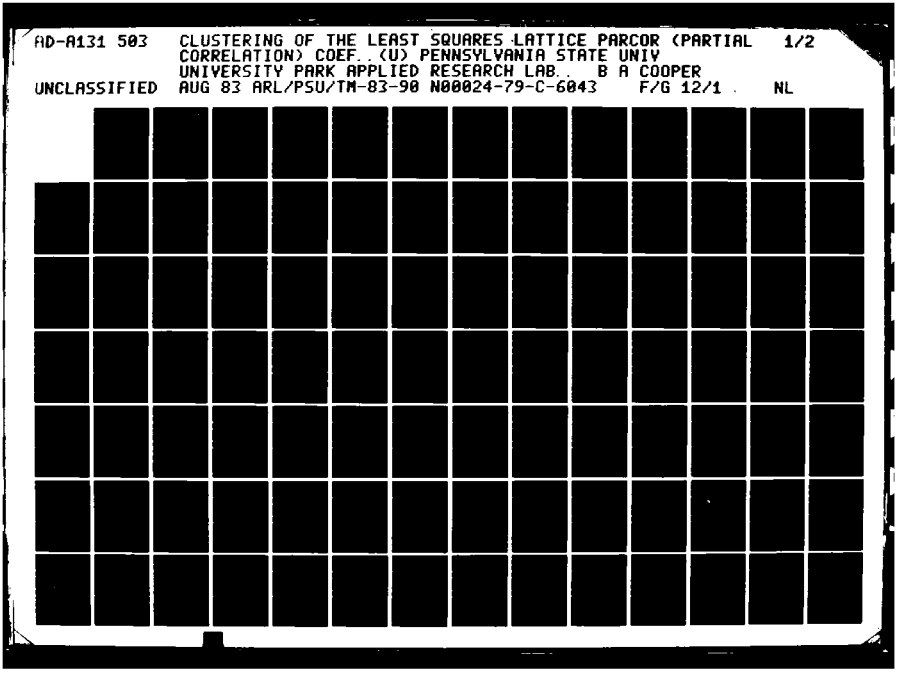
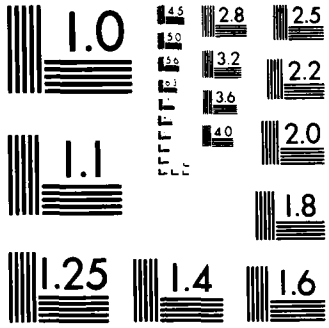

AD-A131 503 CLUSTERING OF THE LEAST SQUARES LATTICE PARCOR (PARTIAL 1/2
CORRELATION) COEF..(U) PENNSYLVANIA STATE UNIV
UNIVERSITY PARK APPLIED RESEARCH LAB.. B A COOPER
UNCLASSIFIED AUG 83 ARL/PSU/TM-83-90 N00024-79-C-6043 F/G 12/1 NL



CLUSTERING OF THE LEAST SQUARES LATTICE PARCOR COEFFICIENTS: A PATTERN-RECOGNITION APPROACH TO STEADY STATE SYNTHETIC VOWEL IDENTIFICATION

Beth Anne Cooper

Technical Memorandum
File No. TM 83-90
June 1, 1983
Contract No. N00024-79-C-6043

Copy No. 5

The Pennsylvania State University
Intercollege Research Programs and Facilities
APPLIED RESEARCH LABORATORY
Post Office Box 30
State College, PA 16801

NAVY DEPARTMENT

NAVAL SEA SYSTEMS COMMAND

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 83-90	2. GOVT ACCESSION NO. AD-A131503	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) CLUSTERING OF THE LEAST SQUARES LATTICE PARCOR COEFFICIENTS: A PATTERN-RECOGNITION APPROACH TO STEADY STATE SYNTHETIC VOWEL IDENTIFICATION		5. TYPE OF REPORT & PERIOD COVERED M.S. Thesis, August 1983 6. PERFORMING ORG. REPORT NUMBER 83-90
7. AUTHOR(s) Beth Anne Cooper		8. CONTRACT OR GRANT NUMBER(s) N00024-79-C-6043
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Pennsylvania State University Applied Research Laboratory, P.O. Box 30 State College, PA 16801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Sea Systems Command Department of the Navy Washington, DC 20362		12. REPORT DATE June 1, 1983 13. NUMBER OF PAGES 122 pp.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified, Unlimited 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited, per NSSC (Naval Sea Systems Command), June 30, 1983

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

thesis, clustering, least, squares, lattice, coefficients

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The partial correlation (PARCOR) coefficients of the least squares lattice filter may be used to conveniently and efficiently represent various types of acoustic signals. Because a stationary time series may be represented by a small number of PARCOR coefficients, the PARCOR coefficients have been widely used as effective pattern recognition parameters for the representation and transmission of information. This thesis establishes the PARCOR coefficients of the least squares lattice filter as efficient and effective pattern recog-

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

nition features for the classification and identification of synthesized steady state vowel-like sounds. The PARCOR coefficient technique is shown to be a much quicker and more computationally efficient method of vowel identification than identification by formant frequencies, which involves the computation of poles and zeros and the back-calculation of formant frequencies and formant bandwidths.

It is well documented in the literature that steady state vowel sounds may be identified and classified according to their formant frequencies. The formant frequencies of each vowel may be shown to cluster together, and the vowel clusters may be separated from one another in the space defined by some subset of the formant frequencies(Barney, 1952; Peterson, 1952; Peterson & Barney, 1952; Potter & Steinberg, 1950).

When a spoken vowel is presented in time series form, utilization of these clustering properties requires a transformation from time domain to frequency domain (formant frequencies) which is quite complicated and computationally expensive. A more efficient vehicle for classifying steady state vowel-like sounds is developed by this author to be the forward PARCOR coefficient K_i^e, i=1,2,...p which arise naturally as intermediate parameters in a pth-order least squares complex adaptive lattice filter (Hodgkiss & Presley, 1981, 1982).

The formant frequency data used in this study are those measured by Peterson and Barney (1952; Barney, 1952; Potter & Steinberg, 1950), obtained through the courtesy of the Bell Laboratories Archives. A time series for each vowel utterance is generated from three formant frequencies using a six-pole IIR digital recursive filter. The time series are then inverse filtered via a six-zero complex adaptive lattice filter (Alexandrou & Hodgkiss, Note 1; Hodgkiss & Presley, 1981, 1982), producing, for each utterance, a set of six PARCOR coefficients.

The PARCOR coefficients produced by the lattice exhibit the same clustering properties as do the formant frequencies; namely, minimum cluster size (average intracluster distance) in two dimensions for all vowels, and maximum cluster separability (intercluster distance) in six dimensions for selected adjacent-vowel pairs. As a combined measure of compactness and separability, the ratio of the sum of average intracluster distance to intercluster distance for each of the adjacent-vowel pairs yielded roughly equivalent results for the formant frequencies and PARCOR coefficients. Graphically, the first two PARCOR coefficients are sufficient for the identification of the first nine vowels, whereas the third PARCOR coefficient is necessary for identification of the tenth vowel, /3/. These results are analogous to those observed for the clustering of formant frequencies.

ABSTRACT

The partial correlation (PARCOR) coefficients of the least squares lattice filter may be used to conveniently and efficiently represent various types of acoustic signals. Because a stationary time series may be represented by a small number of PARCOR coefficients, the PARCOR coefficients have been widely used as effective pattern recognition parameters for the representation and transmission of information. This thesis establishes the PARCOR coefficients of the least squares lattice filter as efficient and effective pattern recognition features for the classification and identification of synthesized steady state vowel-like sounds. The PARCOR coefficient technique is shown to be a much quicker and more computationally efficient method of vowel identification than identification by formant frequencies, which involves the computation of poles and zeros and the back-calculation of formant frequencies and formant bandwidths.

It is well documented in the literature that steady state vowel sounds may be identified and classified according to their formant frequencies. The formant frequencies of each vowel may be shown to cluster together, and the vowel clusters may be separated from one another in the space defined by some subset of the formant frequencies (Barney, 1952; Peterson, 1952; Peterson & Barney, 1952;

Potter & Steinberg, 1950)

When a spoken vowel is presented in time series form, utilization of these clustering properties requires a transformation from time domain to frequency domain (formant frequencies) which is quite complicated and computationally expensive. A more efficient vehicle for classifying steady state vowel-like sounds is developed by this author to be the forward PARCOR coefficients K_i^e, $i=1,2, \ldots p$ which arise naturally as intermediate parameters in a pth-order least squares complex adaptive lattice filter (Hodgkiss & Presley, 1981, 1982).

The formant frequency data used in this study are those measured by Peterson and Barney (1952; Barney, 1952; Potter & Steinberg, 1950), obtained through the courtesy of the Bell Laboratories Archives. A time series for each vowel utterance is generated from three formant frequencies using a six-pole IIR digital recursive filter. The time series are then inverse filtered via a six-zero complex adaptive lattice filter (Alexandrou & Hodgkiss, Note 1; Hodgkiss & Presley, 1981, 1982), producing, for each utterance, a set of six PARCOR coefficients.

The PARCOR coefficients produced by the lattice exhibit the same clustering properties as do the formant frequencies; namely, minimum cluster size (average

intracluster distance) in two dimensions for all vowels, and maximum cluster separability (intercluster distance) in six dimensions for selected adjacent-vowel pairs. As a combined measure of compactness and separability, the ratio of the sum of average intracluster distances to intercluster distance for each of the adjacent-vowel pairs yielded roughly equivalent results for the formant frequencies and PARCOR coefficients. Graphically, the first two PARCOR coefficients are sufficient for the identification of the first nine vowels, whereas the third PARCOR coefficient is necessary for identification of the tenth vowel, /ɜ/. These results are analogous to those observed for the clustering of formant frequencies.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

ACKNOWLEDGMENTS

This research was performed at the Applied Research Laboratory of the Pennsylvania State University under contract with the Naval Sea Systems Command. This support is gratefully acknowledged.

I would like to thank Bell Laboratories, Inc. for the use of the vowel formant frequency data (Note 2). I am especially indebted to Dr. Marcy Goldstein of the Bell Laboratories Archives, who was instrumental in finding the data and supplying me with it. The critical review supplied by Bell Laboratories, Inc. is also very much appreciated.

The least squares lattice software was adapted from a package of unpublished programs supplied by D. Alexandrou and W. S. Hodgkiss (Note 1) of the Scripps Institution of Oceanography, San Diego. Dr. Hodgkiss also supplied extremely helpful background material and notes (Note 3) on the theoretical derivation of the least squares lattice equations (Pack & Satorius, Note 4).

I am also indebted to the following colleagues at the Applied Research Laboratory, Pennsylvania State University: John Sacha for his valuable programming assistance and critical review, Guy Sohie for many helpful discussions and a critical review, and Regina Kicera for use of her statistical analysis software. Also, I appreciate the fine

work done on my behalf by the staffs of the ARL Library and the Editorial Department.

Most importantly, I would like to express my sincere gratitude to the members of my thesis committee, whose guidance and insight have benefitted me immensely: Dr. Harvey Gilbert and Dr. John Lewis, whom I asked to serve in this capacity because they have been inspiring, demanding, dedicated and fair as classroom teachers; and my thesis adviser, Dr. Leon Sibul, who has been especially patient with me and very supportive of my interests.

CHAPTER I

INTRODUCTION

The partial correlation (PARCOR) coefficients of the least squares lattice may be used to conveniently and efficiently represent various types of acoustic signals. Because a stationary time series may be represented by a small number of PARCOR coefficients, the PARCOR coefficients have been widely used as effective pattern recognition parameters for the representation and transmission of information. This thesis establishes the PARCOR coefficients of the least squares lattice as efficient and effective pattern recognition features for the classification and identification of synthesized steady state vowel-like sounds. The PARCOR coefficient technique is shown to be a much quicker and more computationally efficient method of vowel identification than identification by formant frequencies, which involves the computation of poles and zeros and the back-calculation of formant frequencies and formant bandwidths. Even though this thesis addresses identification of vowels, it is clear that the PARCOR coefficients may be used to identify characteristics of other acoustic and electromagnetic signals.

It is well documented in the literature that steady state vowel sounds may be identified and classified

according to their formant frequencies. The formant frequencies of each vowel may be shown to cluster together, and the vowel clusters may be separated from one another in the space defined by some subset of the formant frequencies (Barney, 1952; Peterson, 1952; Peterson & Barney, 1952; Potter & Steinberg, 1950)

When a spoken vowel is presented in time series form, utilization of these clustering properties requires a transformation from time domain to frequency domain (formant frequencies) which is quite complicated and computationally expensive. A more efficient vehicle for classifying steady state vowel-like sounds is developed by this author to be the forward PARCOR coefficients K_i^e, $i=1,2, \ldots p$ which arise naturally as intermediate parameters in a pth-order least squares complex adaptive lattice filter (Hodgkiss & Presley, 1981, 1982).

The specific purpose of this research is to establish the value of the PARCOR coefficients as efficient and effective pattern recognition features for the classification and identification of (synthesized) steady state vowel-like sounds. Specifically, the intent of this thesis is to show that for steady state vowel-like utterances, the PARCOR coefficients of each vowel will cluster together, and that the ten English vowels may be separated from one another in the space defined by some

subset of the PARCOR coefficients in much the same way as the formant frequencies cluster and separate the vowels. In other words, the inverse filtering procedure may be considered as a change of variable (Turner, 1982); the PARCOR coefficients behave in this manner when the related formant frequencies themselves exhibit a clustering behavior.

A time series for each vowel utterance is generated from three formant frequencies using a six-pole IIR recursive digital filter. The formant frequency data are those measured by Peterson and Barney (1952; Barney, 1952; Potter & Steinberg, 1950), obtained through the courtesy of the Bell Laboratories Archives. A six-zero complex adaptive least squares lattice filter is used as an inverse filter on each time series, producing, for each, a set of six PARCOR coefficients.

Considerable motivation exists for the development of a system identification technique which does not require the calculation of formant frequencies and bandwidths from a time series. Specifically, in the field of speech processing (Markel, 1972, 1973; McCandless, 1974; Wakita & Kasuya, 1977), researchers have commonly calculated the coefficients, $\hat{a}_j$, of the denominator polynomial from the transfer function of an inverse filter and then obtained the formant frequencies from the roots of the polynomial. As

stated previously, this is a complicated calculation. For applications where frequencies of the formants are not important but desired for vowel identification clustering, a change of clustering variable to the PARCOR coefficients would eliminate the expensive and complicated computation.

Vowel Identification by Formant Frequency

During voiced speech, when the vocal tract is excited by the glottal source, the spectral peaks which occur are referred to as the formants of the particular speech sound. For the majority of male speakers the first three formants lie in the ranges 150-850 Hz., 500-2500 Hz., and 1700-3500 Hz.. Formants for women and children are higher in frequency, due, in part, to the smaller size of their vocal mechanisms (Fant, 1956).

Some speech sounds, such as steady state vowels, may be identified or characterized by their formant frequencies and the bandwidths and levels of those formant frequencies. When different speakers speak one of the vowels, the utterances are different for each speaker. In the perceptual space defined by the frequencies of the formants (which is referred to as the formant space), these differences manifest themselves, for each vowel, as a cluster of points around some average value. In the speech literature, the first three formants are used widely for

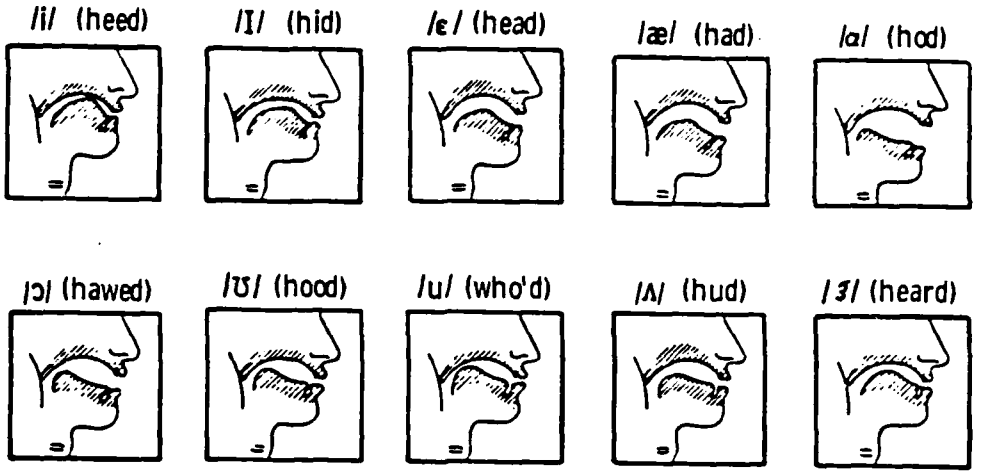
adequate vowel identification, (Peterson, 1952; Peterson & Barney, 1952; Potter & Steinberg, 1950), although considerable evidence has been offered in the literature both for and against the necessity of more formants, formant bandwidths, and/or formant amplitudes (Peterson, 1952; Peterson, 1961; Bernstein, 1981; Potter & Steinberg, 1950), and fundamental frequencies of excitation (Foulkes, 1961; Peterson, 1961) for vowel identification. A popular three-dimensional mechanism for vowel identification is a plot of the first three formant frequencies F_1, F_2, F_3, in the perceptual coordinate space defined by the axes F1, F2, and F3. A base 10 logarithmic scale is usually used to account for the nonlinearities of the ear. The spatial position of the articulators and the vocal mechanism at any point in time directly affect the frequency position of the resonances of the vocal tract by changing the relative sizes of different parts of the tract.

The location of a vowel in the "formant space" defined by F1 and F2 corresopnds to the spatial location of the tongue hump in the two-dimensional representation of the oral cavity for that vowel. In other words, the formant frequency space classification seems to have a physical significance. Figure 1 shows the configuration of the articulators for the ten English vowels (adapted from Potter, Kopp, & Kopp, 1966, chap. 12) and their

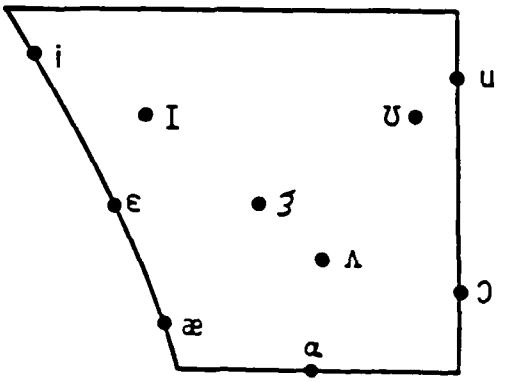
corresponding relative tongue hump positions in the vowel quadrilateral (adapted from Denes & Pinson, 1963, p. 55). If the vowels are classified based on these tongue hump positions; (i.e., /i/, as in the word "heed" is a high, front vowel, whereas /ɔ/, as in the word "hawed" is a low back vowel), the classifications are similar to those obtained using the formant space plots. The Peterson and Barney study is discussed and results of graphical and quantitative analyses of the formant frequency data are presented in Chapter II.

Acoustic Tube Vocal Tract Models

There are many applications where the PARCOR coefficients have a direct physical relationship to sound generation mechanisms. One such case is the generation of speech sounds by the human vocal system. The vocal tract is often modeled as an acoustic tube (Dunn, 1961; Flanagan, 1972; Markel & Gray, 1976, chap. 4; Wakita, 1973a, 1973b, 1979; Wakita & Gray, 1975) excited by either a voiced or unvoiced source somewhere along its length, with appropriate boundary conditions which depend on the circumstances of phonation. The PARCOR coefficients of the lattice structure have a direct physical relationship to the reflections between sections of the acoustic tube. Specifically, the area ratio between successive sections is defined



(a)



(b)

Figure 1. (a) Schematic vocal tract profiles for the production of English vowels. (Adapted from Potter, Kopp, & Kopp, 1966, chap. 12) (b) Tongue hump positions for the ten English vowels. (Adapted from Denes & Pinson, 1963, p. 55)

(Wakita, 1979, p. 281) as

$$\frac{A_{j+1}}{A_j} = \frac{1-Kj}{1+Kj} \qquad j=1, 2, \ldots p.$$

where the A_j are the areas of the sections of the acoustic tube and the Kj are the PARCOR coefficients for a pth-order model. Wakita (1973b) suggested that the area function of the acoustic tube (determined from a ladder implementation of the linear prediction technique) could be used to detect obstructions in the vocal tract. If the tube is considered to be lossless, of length L, constant cross-sectional area, and closed at the vocal fold end while open at the lip end, resonances occur at the frequencies corresponding to $L=n\lambda/4$; $n=1, 2, \ldots$ where λ is the wavelength of the fundamental resonant frequency.

Autoregressive Model for Vowel Representation

The acoustic tube model of the vocal tract lends itself directly to a mathematical model for a spoken vowel sound. Non-nasal vowels have been modeled widely in the speech literature as autoregressive (AR) processes of order p, generated by passing a white input time series, v(n), through a pth-order all pole filter with transfer function

$$H(z)= \frac{1}{A(z)} = \frac{Z(z)}{V(z)}$$

where $A(z)=1-\sum_{j=1}^{p} a_j z^{-j}$ and the output process,

$$z(n)= \sum_{j=1}^{p} a_j z(n-j)+v(n).$$

The inverse of a pth-order AR process is a pth-order moving average (MA) process, generated by passing the input, v(n), through a pth-order all zero filter with transfer function

$$H(z) = B(z) = \frac{Z(z)}{V(z)}$$

where $B(z)= 1- \sum_{k=1}^{p} b_k z^{-k}$ and $z(n)= \sum_{k=1}^{p} b_k\ v(n-k)$.

The least squares lattice filter used in this study is a feed-forward MA (all-zero) lattice. This inverse lattice filter produces a whitened output spectrum by placing a zero at the location of each pole in the input spectrum. Optimal whitening is obtained when the following conditions are satisfied (because of the one to one correspondence of zeros to poles): 1) AR input processes are used as the lattice input, and 2) the order of the lattice filter chosen to be equal to the order of the input AR process. The fitting of an AR model to a time series is equivalent to the method of maximum entropy spectral analysis (Burg, 1967; Macina, 1981; Papoulis, 1981; Parzen, 1974; Ulrych & Bishop, 1975). The maximum entropy technique assumes maximum uncertainty with respect to the unknown information about the signal (outside the sampling interval).

Because the aim of this research is to illustrate the pattern recognition capabilities of the PARCOR coefficients for data with known formant frequencies (data which had previously been shown to cluster in the frequency domain), the vowels are modeled as sixth-order AR processes and the lattice is specified to be sixth-order to maximize the accuracy of the PARCOR coefficients. The modeling of a vowel sound as a sixth-order AR process is a gross over-simplification in terms of speech production. However, it is a necessary one if the fundamental intent of the research is to be respected. Generation of vowel utterances from formant frequencies is discussed in Chapter III.

The Linear Prediction Problem

The inverse filter is determined through the use of the linear prediction technique. This technique has been used widely in the speech field as well as in geophysics and neurophysics (Landers & Lacoss, 1977; Makhoul, 1975; Wood & Treitel, 1975) for time series modeling. The linear predictive technique estimates the properties of a signal by modeling a sample as a linear combination of past samples and minimizing some form of the error between the actual and predicted samples. The most common implementations of the linear predictive technique have been the autocorrelation and covariance methods (Makhoul, 1975; Markel & Gray, 1976, chap. 9; Rabiner & Schafer, 1978).

Lattice Method of Linear Prediction

The linear prediction technique may also be implemented as a lattice algorithm, which is recursive in time and order. A lattice structure has several advantages over the autocorrelation and covariance methods. Most importantly, the physical structure of the lattice filter is composed of cascaded filter stages; a pth-order lattice filter may be decomposed into all filters of up to and including pth-order. The pth-order lattice structure simultaneously generates outputs of all lesser order filters. Lattice models are naturally related to physical models such as the scattering and propagation of waves in a stratified medium (Friedlander, 1980). The PARCOR coefficients of the lattice are also related to the reflections between layers of the medium being modeled. This type of physical meaning is not apparent for the polynomial filter coefficients, $\hat{a}_j$, which are obtained from the PARCOR coefficients by a nonlinear recursion. The lattice model is directly applicable to the study of transmission theory, seismic signal processing, and underwater sound propagation as well as speech communication. For instance, geophysicists have used ladder structures in their study of structural features of the earth's subsurface (Burg, 1967; Robinson & Treitel, 1980; Wiggins & Robinson, 1965; Ulrych & Bishop, 1975). The lattice method of linear prediction was first presented by

Itakura & Saito (1971) and is well known in the speech field as an analysis tool (Atal & Hanauer, 1971; Flanagan, 1972; Makhoul, 1975). Also, hardware implementations of lattice filters have been successfully marketed as effective synthesis devices for the compression and transmission of speech signals. The Texas Instruments' "Speak & Spell" game is an example of this technology. Researchers in the speech field have used the lattice based on an acoustic tube model of the vocal tract (Markel & Gray 1976, chap. 4; Wakita, 1973a, 1973b; Wakita & Gray 1975). In addition to the obvious physical significance of the lattice structure, other advantageous features of the lattice method of linear prediction include a recursive-in-time implementation (rather than block processing), faster convergence, insensitivity to eigenvalue spread, better numerical behavior, robustness, and insensitivity to roundoff noise (Friedlander, 1982a; Lee, Morf & Friedlander, 1981; Makhoul, 1978; Markel & Gray, 1976, chap. 9). Lattice filters have found application in the fields of noise cancelling, channel equalization, seismic signal processing, speech processing, system identification, frequency tracking, spectral estimation, and spectrum prewhitening (Friedlander, 1982b; Hodgkiss & Alexandrou, 1983; Hodgkiss & Presley, 1981, 1982; Lee, Morf, & Friedlander, 1981; Satorius & Alexander, 1979; Satorius & Pack, 1981; Satorius & Shensa, 1980a).

The least squares lattice. The least squares lattice recursions were first obtained by an algebraic approach (Morf, Lee, Nickolls & Vieira, 1977; Morf, Vieira, & Lee, 1977). The least squares lattice (LSL) structures proposed by Morf et. al are more efficient numerically than the gradient lattice algorithms (Griffiths, 1975, 1977, 1978), requiring only O(p) operations per time update, where p is the order of the filter. The least squares lattice structures are known for their good stability properties, rapid startup, excellent convergence properties and fast parameter tracking capabilities (Hodgkiss & Presley, 1981; Lee, 1980; Lee, Morf, & Friedlander, 1981; Satorius & Pack, 1981; Satorius & Shensa, 1980b). These advantages are a direct result of two lattice parameters which account for the algorithmic differences between the gradient and LSL formulations: an exponential weighting parameter, $(1-\alpha_{CLSL})$, and a Gaussian likelihood step size parameter, $\gamma_{i-2}(n-1)$. These parameters are discussed in Chapter IV. Certain assumptions must be made about the behavior of the waveform outside the time interval of observation. The lattice structure used in this research prewindows the data, i.e., it assumes that $z(n)=0$ for all $n<0$. The LSL Algorithms deveolped by Morf et. al were adapted for complex data by Hodgkiss and Presley (1981, 1982) and then programmed in FORTRAN by Alexandrou and Hodgkiss (Note 1). A section of this program was incorporated into this

author's computer simulation for the inverse filter. The inverse filter is discussed in Chapter IV.

CHAPTER II

VOWEL IDENTIFICATION

The concept of formant space clustering is the result of work done at Bell Laboratories during the years 1947-1951. A study of sustained vowels was undertaken by G. E. Peterson, H. L. Barney, R. K. Potter, J. C. Steinberg and others to investigate the relationship between spoken and perceived vowels and their acoustical correlates (Barney, Note 2; Peterson & Barney, 1952; Potter & Steinberg, 1950). The vowels used were the ten English vowels in a consonant-vowel-consonant (CVC) context with /h/ as the first consonant. For the Bell Laboratories studies, a total of 76 speakers including 33 men, 28 women, and 15 children each recorded two lists of 10 words (each word contained a vowel in the context /h_d/). The vowels and corresponding symbols and words adapted from those of the International Phonetic Alphabet (IPA) are presented in Table 1. The majority of the male speakers spoke General American English. The words were recorded and played to a group of 70 listeners who identified the vowel in each of the words. The vowel portion of each CVC utterance was also analyzed with a sound spectrograph to determine the frequency positions of the first three formants. Since the results of the listening tests showed the effects of the diverse dialectal backgrounds of the listeners, a sub-group

Table 1

Vowel Symbols and Corresponding CVC Test Utterances

Used in the Bell Laboratories Study

Symbol[a]	Data points	CVC syllable	Utterance
/i/	65	/hid/	heed
/I/	51	/hId/	hid
/ɛ/	35	/hɛd/	head
/æ/	56	/hæd/	had
/ɑ/	37	/hɑd/	hod
/ɔ/	45	/hɔd/	hawed
/U/	55	/hUd/	hood
/u/	60	/hud/	who'd
/ʌ/	37	/hʌd/	hud
/ɝ/	65	/hɝd/	heard

[a]Symbols are those of the International Phonetic Alphabet.

of 26 observers with similar characteristics was chosen (Barney, Note 2). Of the 1520 vowels presented, 1203 were identified unanimously by the 26 observers.

When the first and second formants of the vowels are plotted against each other, the vowels appear in essentially the same positions as they do in the vowel quadrilateral (Peterson & Barney, 1952). Vowels may be separated on the basis of their locations in the space defined by the first two formants, except for the vowel /ɜ/, identified by its third formant, which is lower than that of the other vowels (Potter & Steinberg, 1950). The distribution is continuous in the F1-F2 plane in going from vowel to vowel; the overlap between vowels is characteristic of the differences in the way various individuals articulate and pronounce the vowels (Peterson & Barney, 1952). The distributions for each vowel tend to be elongated, elliptical areas along lines which pass through the origin, indicating that although formant ratios are not exactly constant, they do tend to be helpful for the identification of some vowels (Potter & Steinberg, 1950).

Analysis of the Formant Frequency Data

Fundamental and formant frequency data for these 1203 utterances were obtained through the courtesy of Bell Laboratories Archives for use in this study. Because Potter

and Steinberg (1950) determined that formant frequency positions for a man's voice differed from those of a woman's for the same vowel, the present author has restricted this study to data from male utterances to control for fundamental frequency. This author has plotted the vowel data used for this study in the manner of Peterson and Barney. Data for four of the utterances are eliminated from the study because they occurred outside the limits of three standard deviations for a particular vowel. The remaining number of data points per vowel are listed in Table 1. The base 10 logarithms of formant frequencies are normalized so that the frequency range of the entire set of data falls within the interval [0,1]. Table 2 presents the ranges of normalized formant frequencies. The widest range is spanned by F2, the second widest by F1. This is consistent with the recognition of F2 in the literature (Potter, Kopp, & Kopp, 1966, pp. 74-75) as a primary feature of voiced speech, especially the movement of F2 in identifying dipthongs (as in "how", "hoe", "hay", "high", "hoist").

Graphical Analysis

Figures 2 and 3 present the vowel clusters for all ten vowels in the F1-F2 and F1-F3 planes. The same clusters are presented separately by vowel in Figures 4-13 for the F1-F2 plane, and the cluster for the vowel /ɜ/ in the F1-F3 plane is shown in Figure 14. Each data point appears as the

Table 2

Normalized and Un-normalized Ranges for Formant Frequencies and PARCOR Coefficients

Cluster variable	Minimum	Maximum
Un-normalized values		
Formant frequencies (1-3)	190	3400
PARCOR coefficients (1-6)	-0.9674	0.9971
Values normalized to [0,1]		
Formant frequencies		
log F_1	0.0000	0.5235
log F_2	0.3747	0.9201
log F_3	0.6924	1.000
PARCOR coefficients		
K1	0.0000	0.6484
K2	0.3020	1.000
K3	0.0320	0.4346
K4	0.3071	0.9425
K5	0.2027	0.6200
K6	0.8858	0.9303

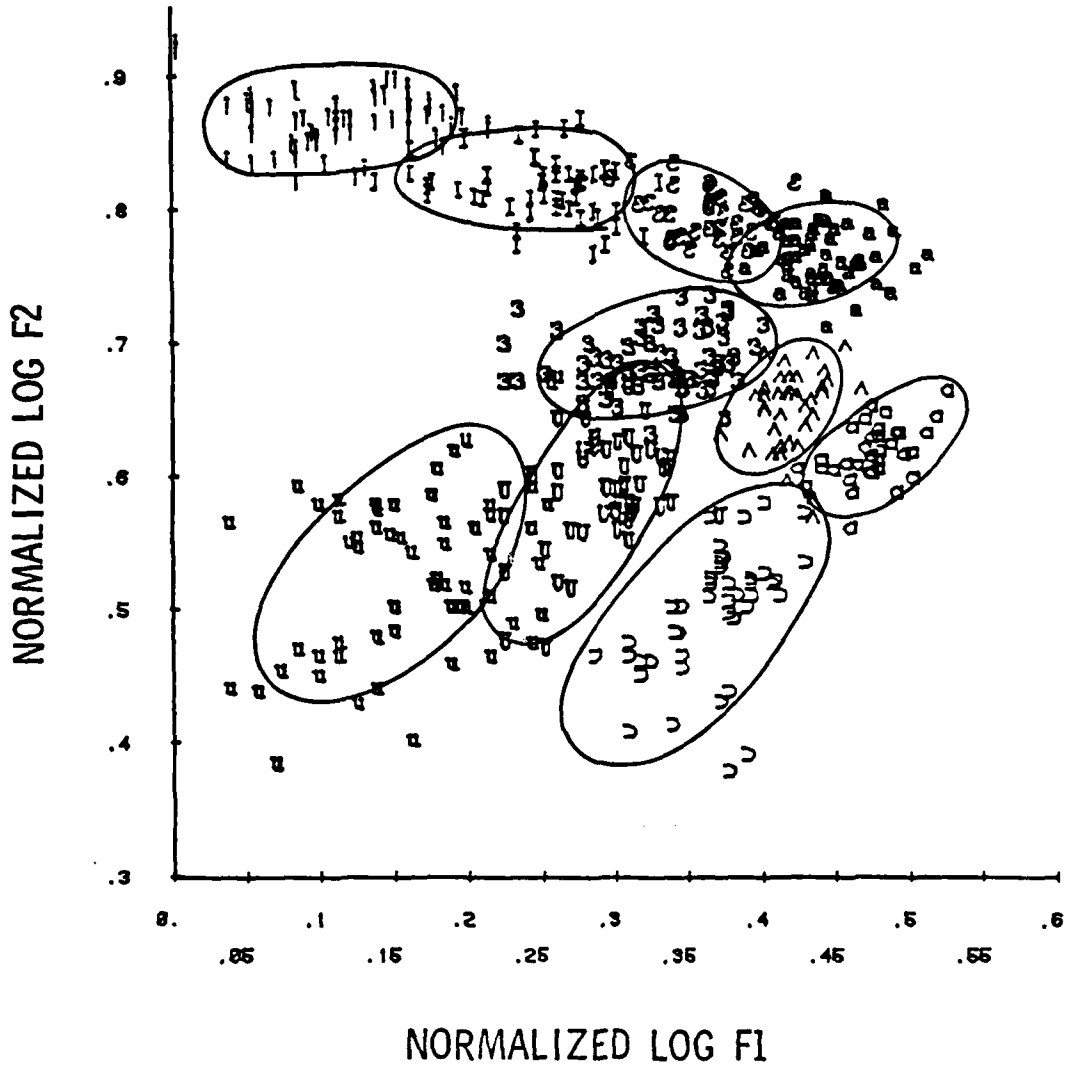


Figure 2. Clustering of the ten English vowels in the F1-F2 plane.

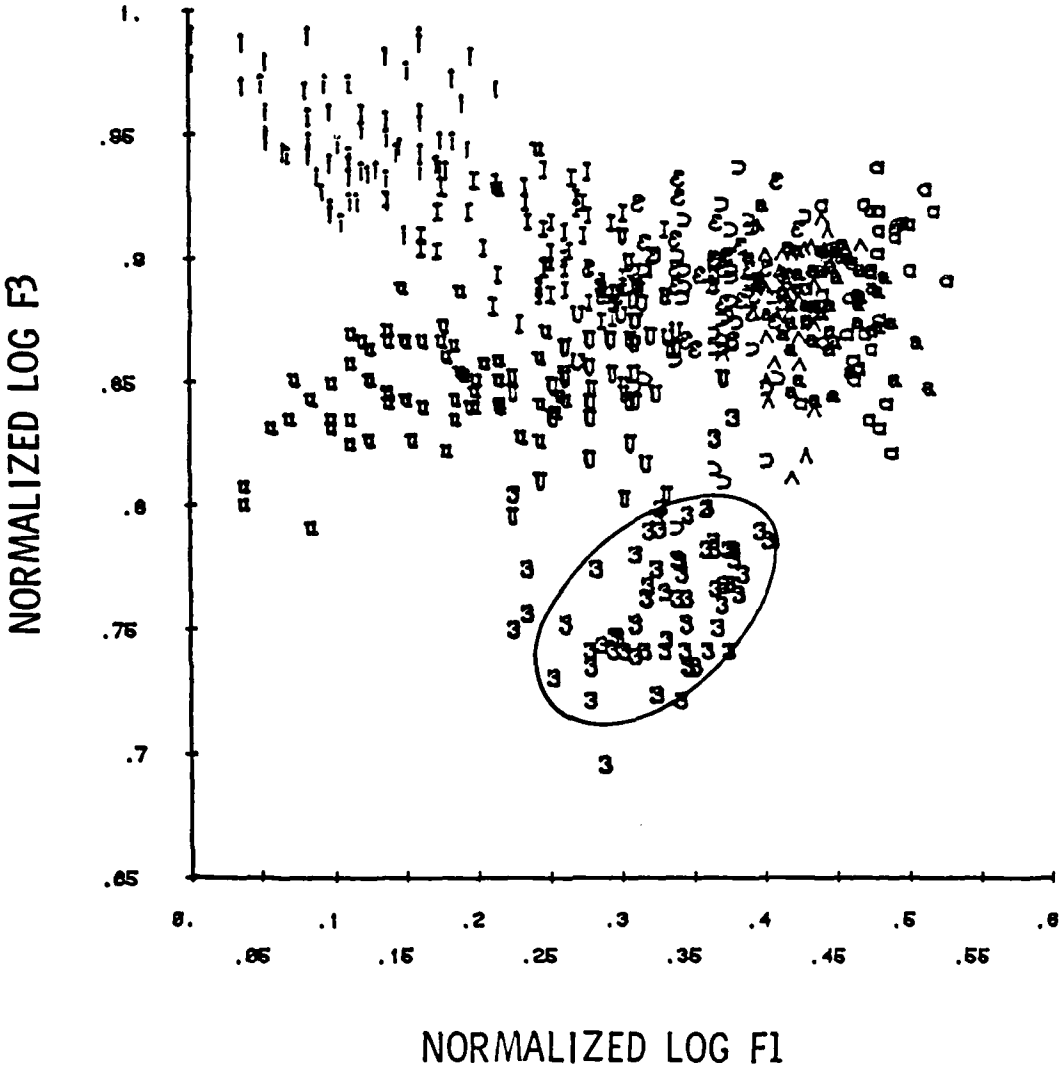


Figure 3. Clustering of the ten English vowels in the F1-F3 plane.

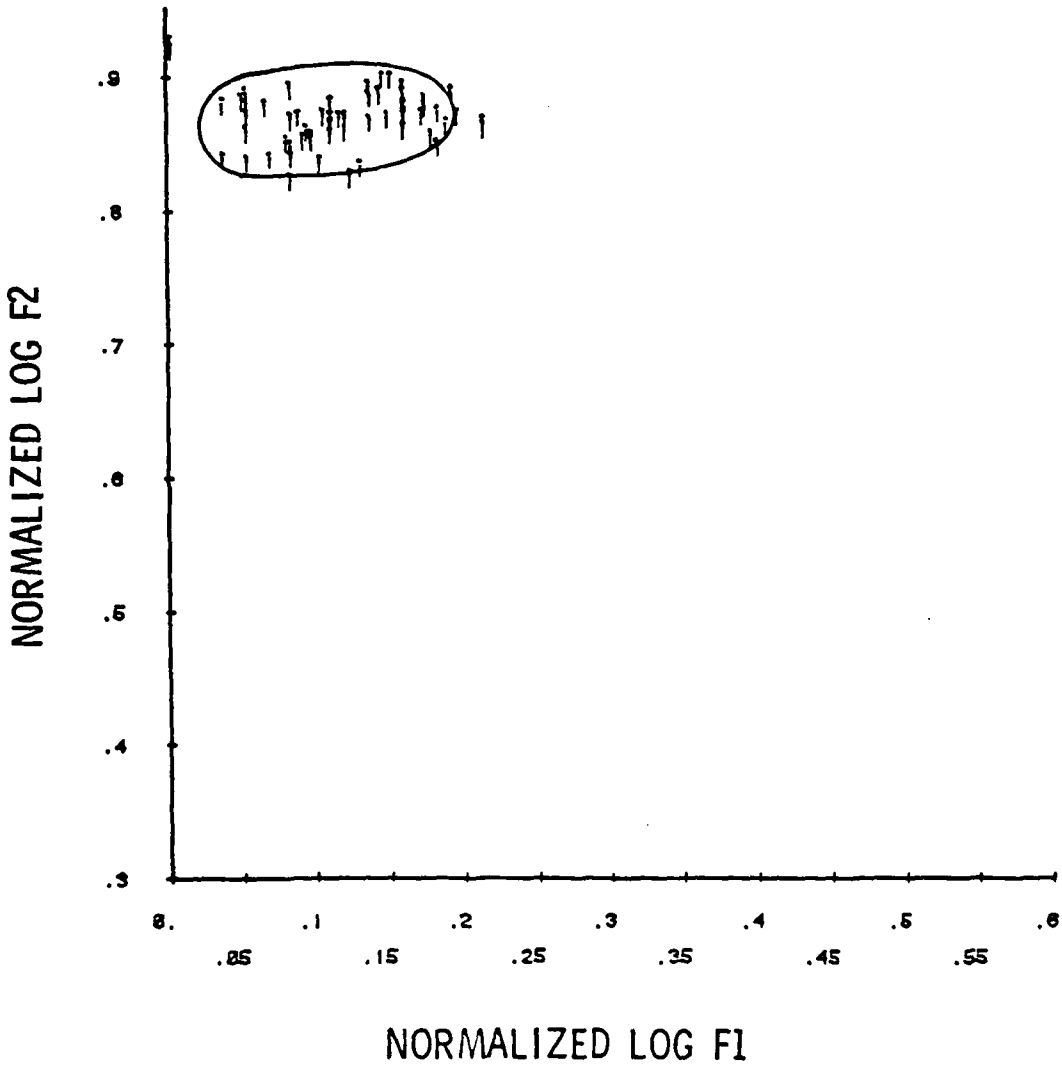


Figure 4. Clustering of the vowel /i/ in the F1-F2 plane.

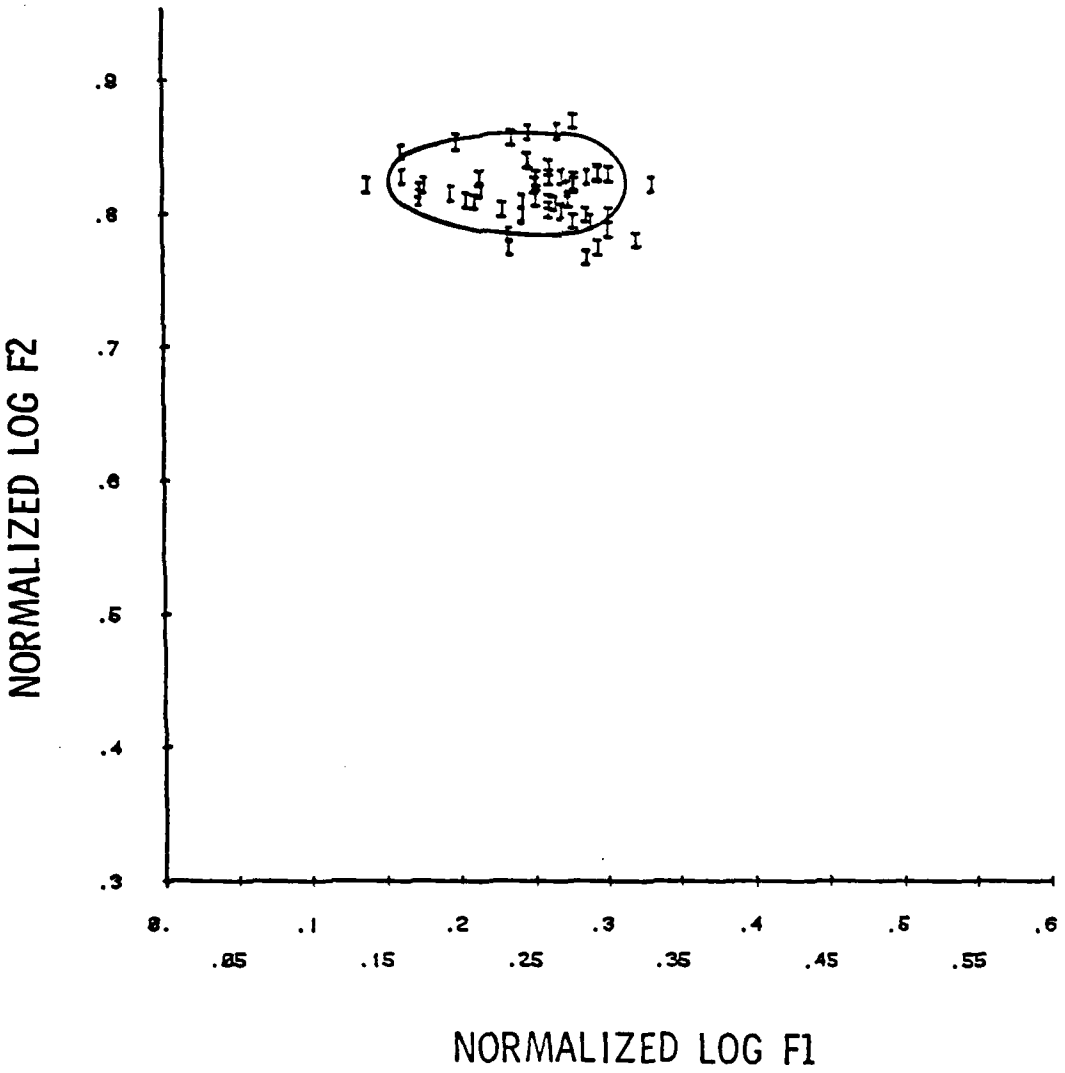


Figure 5. Clustering of the vowel /I/ in the F1-F2 plane.

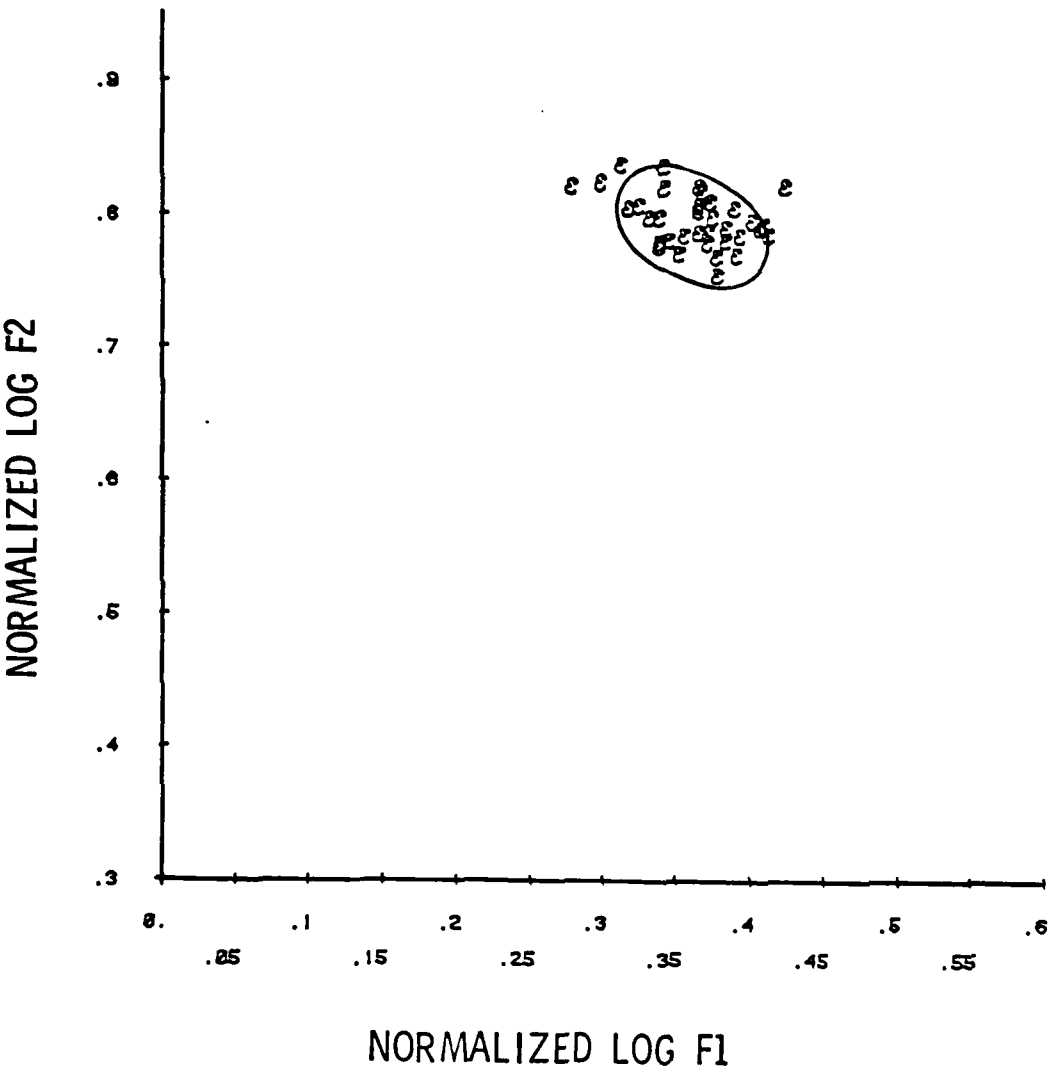


Figure 6. Clustering of the vowel /ɛ/ in the F1-F2 plane.

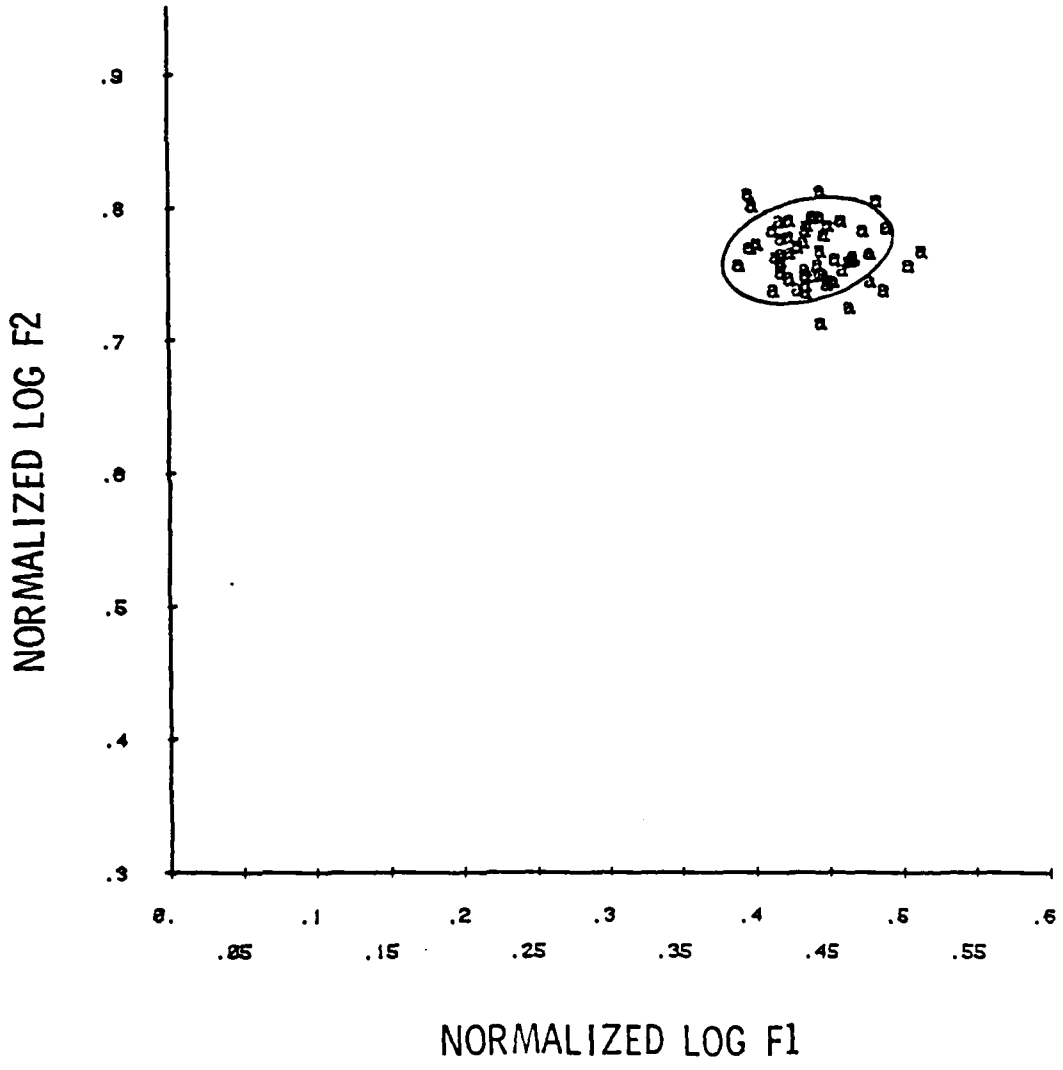


Figure 7. Clustering of the vowel /æ/ in the F1-F2 plane.

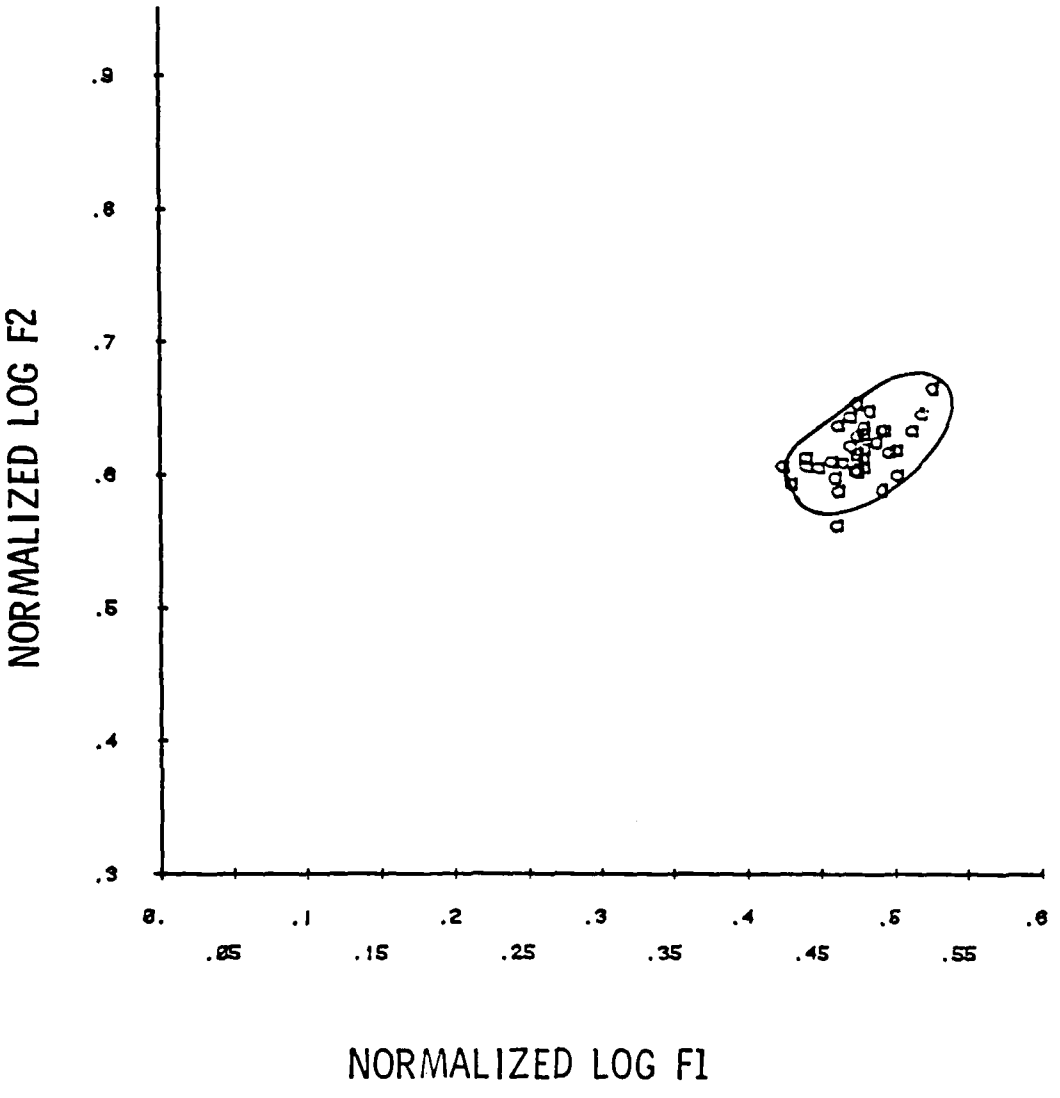


Figure 8. Clustering of the vowel /ɑ/ in the F1-F2 plane.

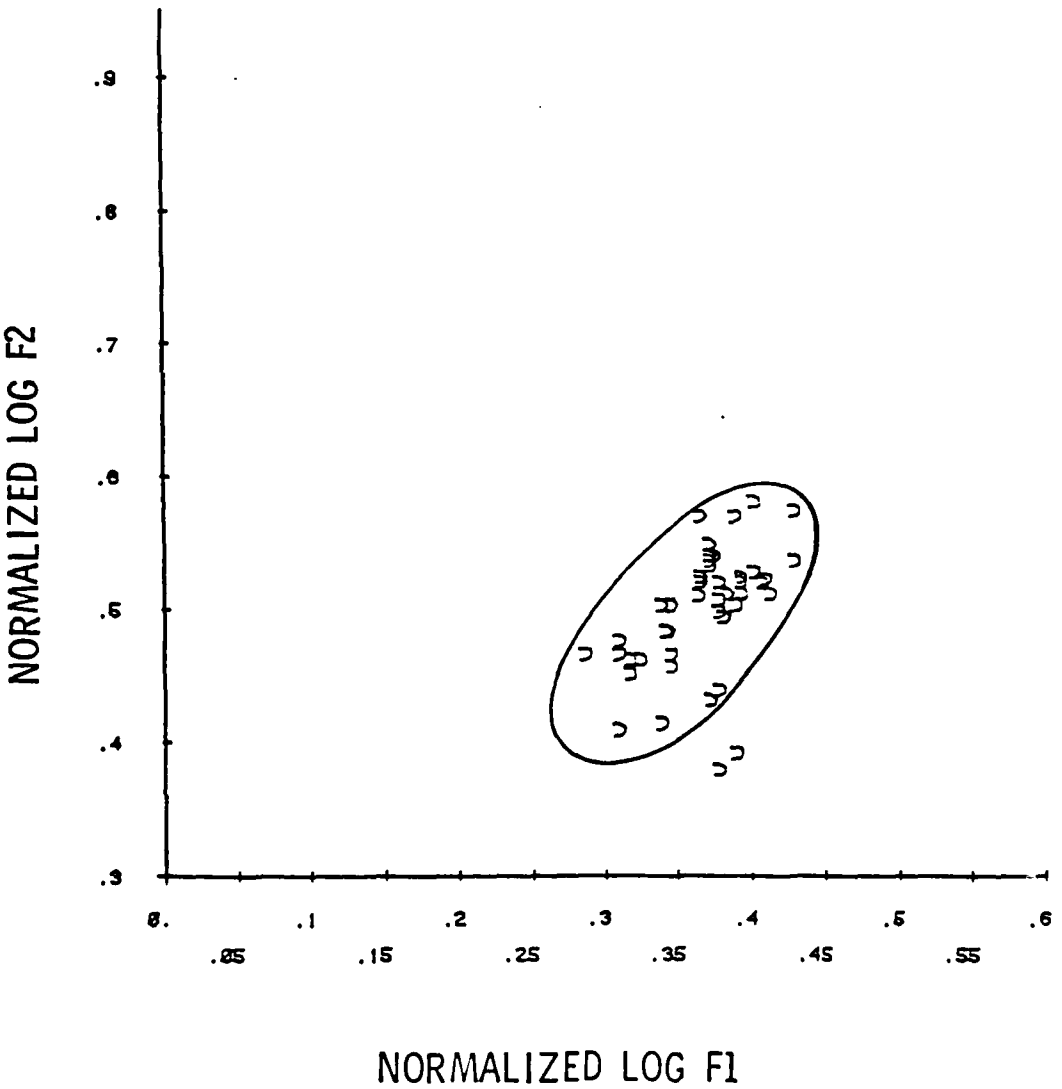


Figure 9. Clustering of the vowel /ɔ/ in the F1-F2 plane.

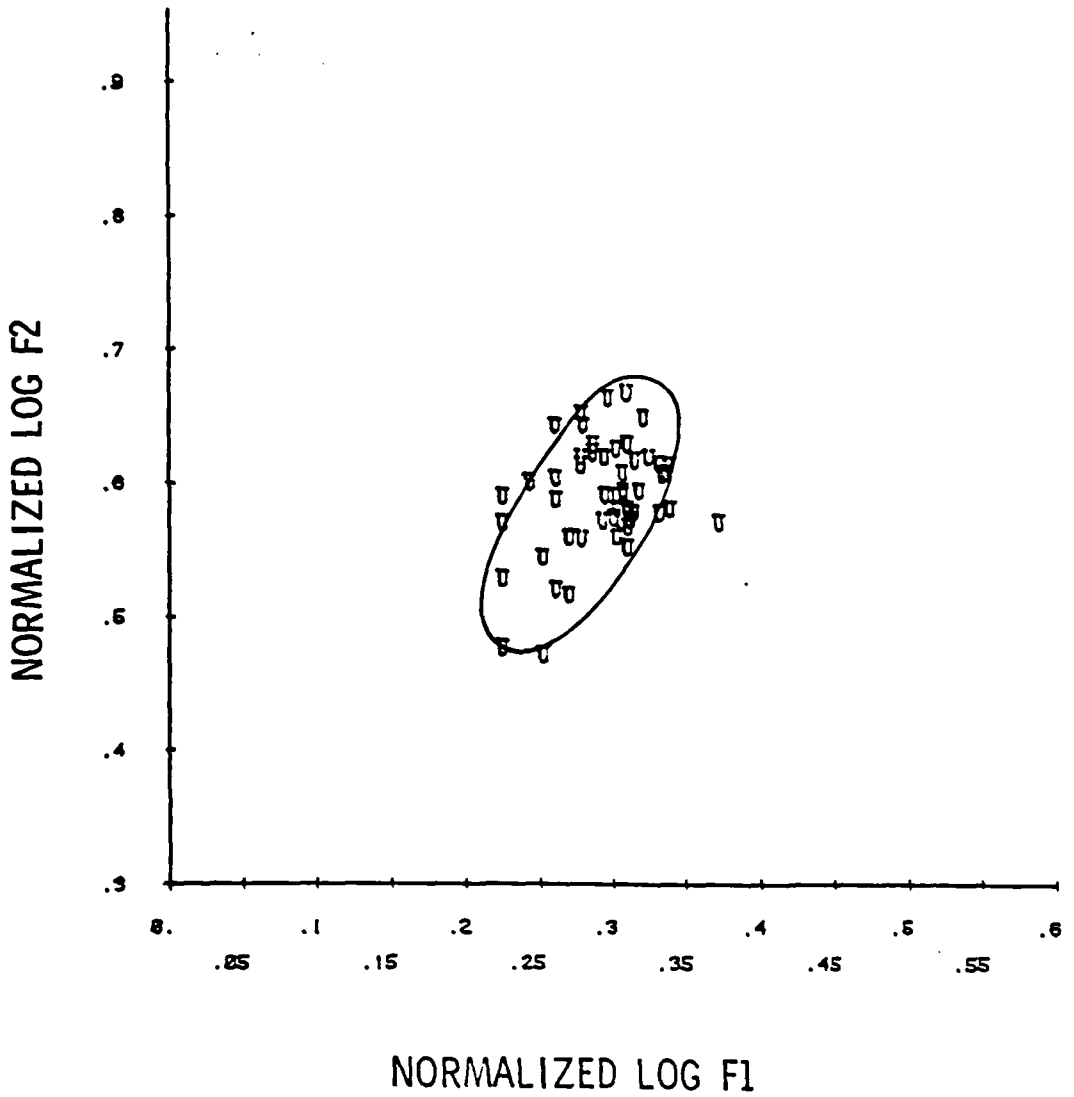


Figure 10. Clustering of the vowel /U/ in the F1-F2 plane.

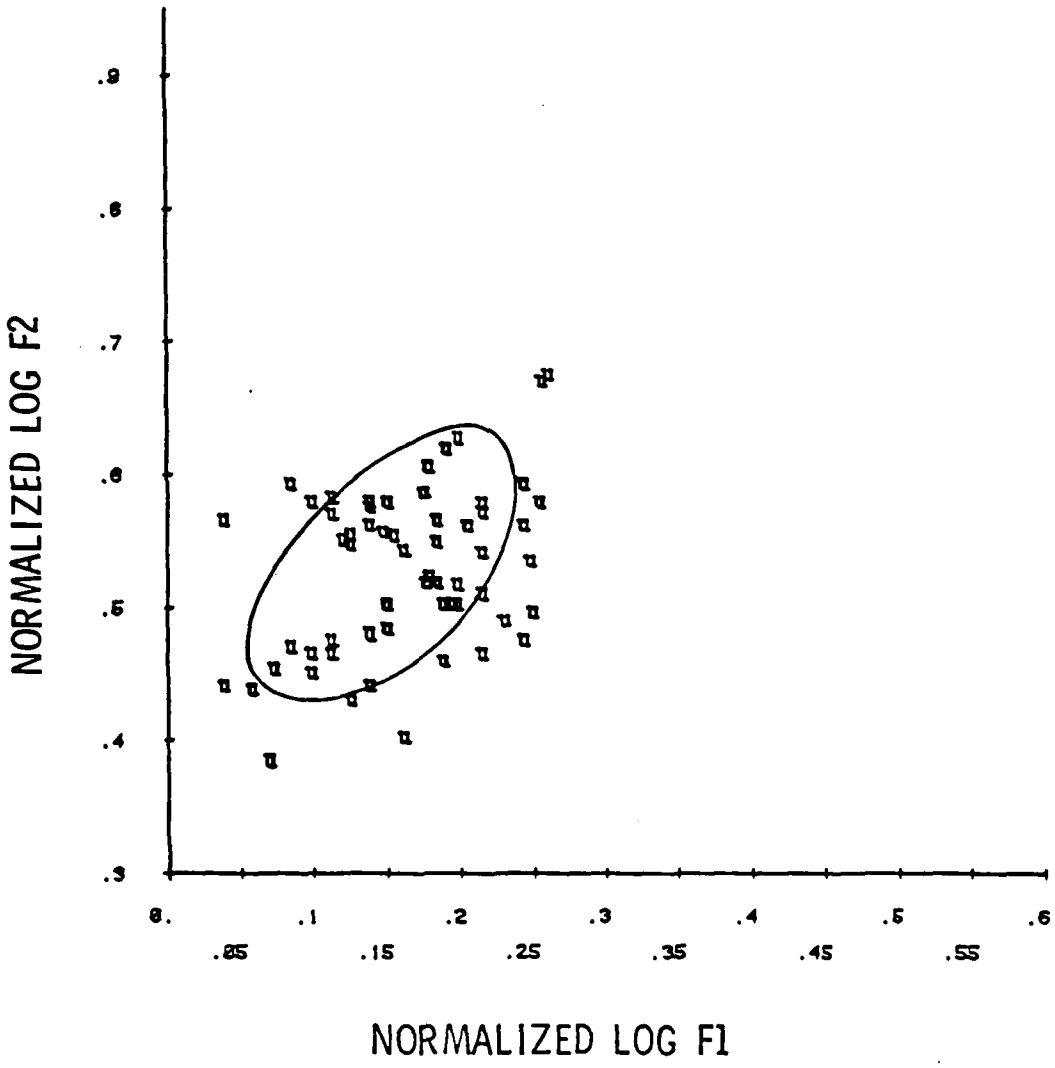


Figure 11. Clustering of the vowel /u/ in the F1-F2 plane.

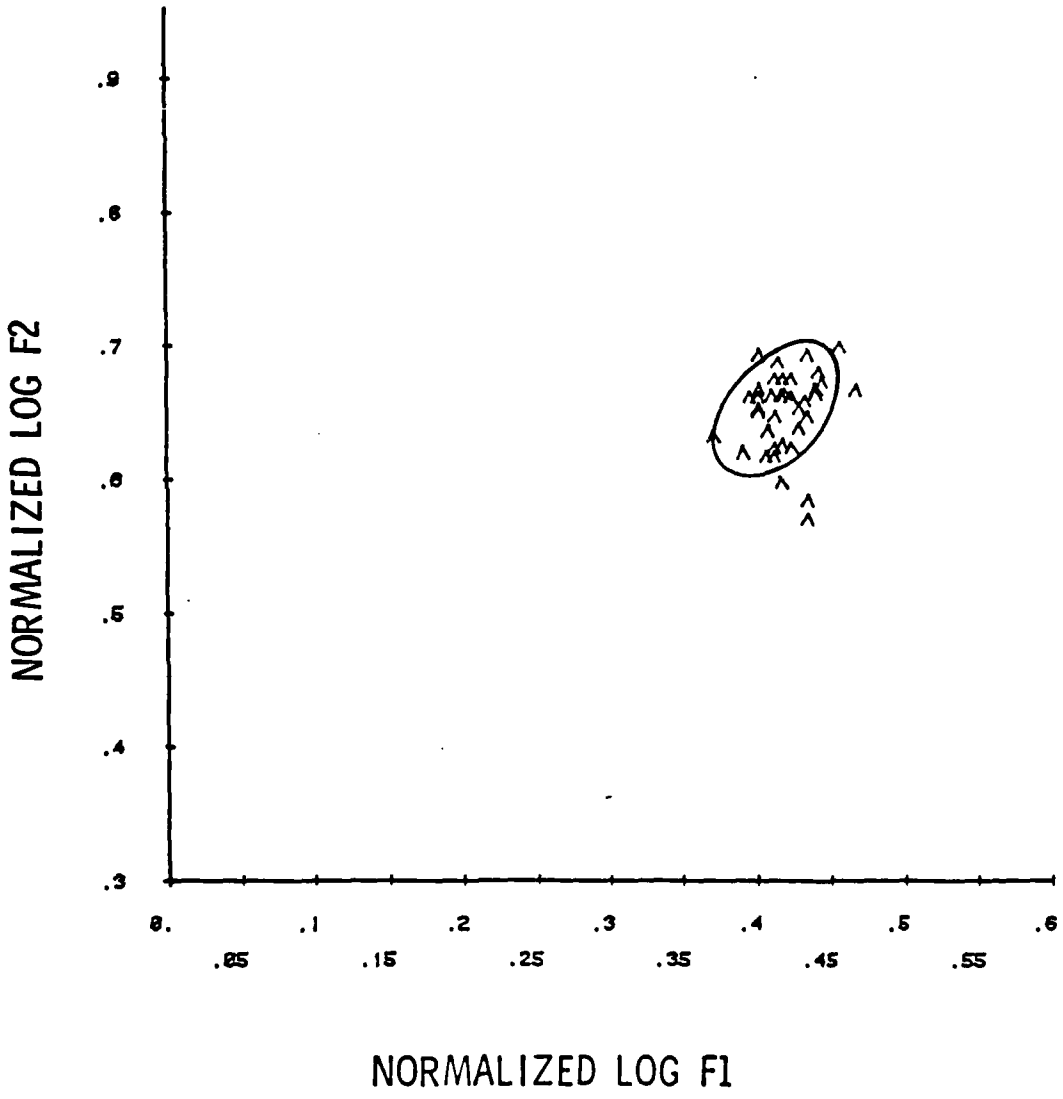


Figure 12. Clustering of the vowel /Λ/ in the F1-F2 plane.

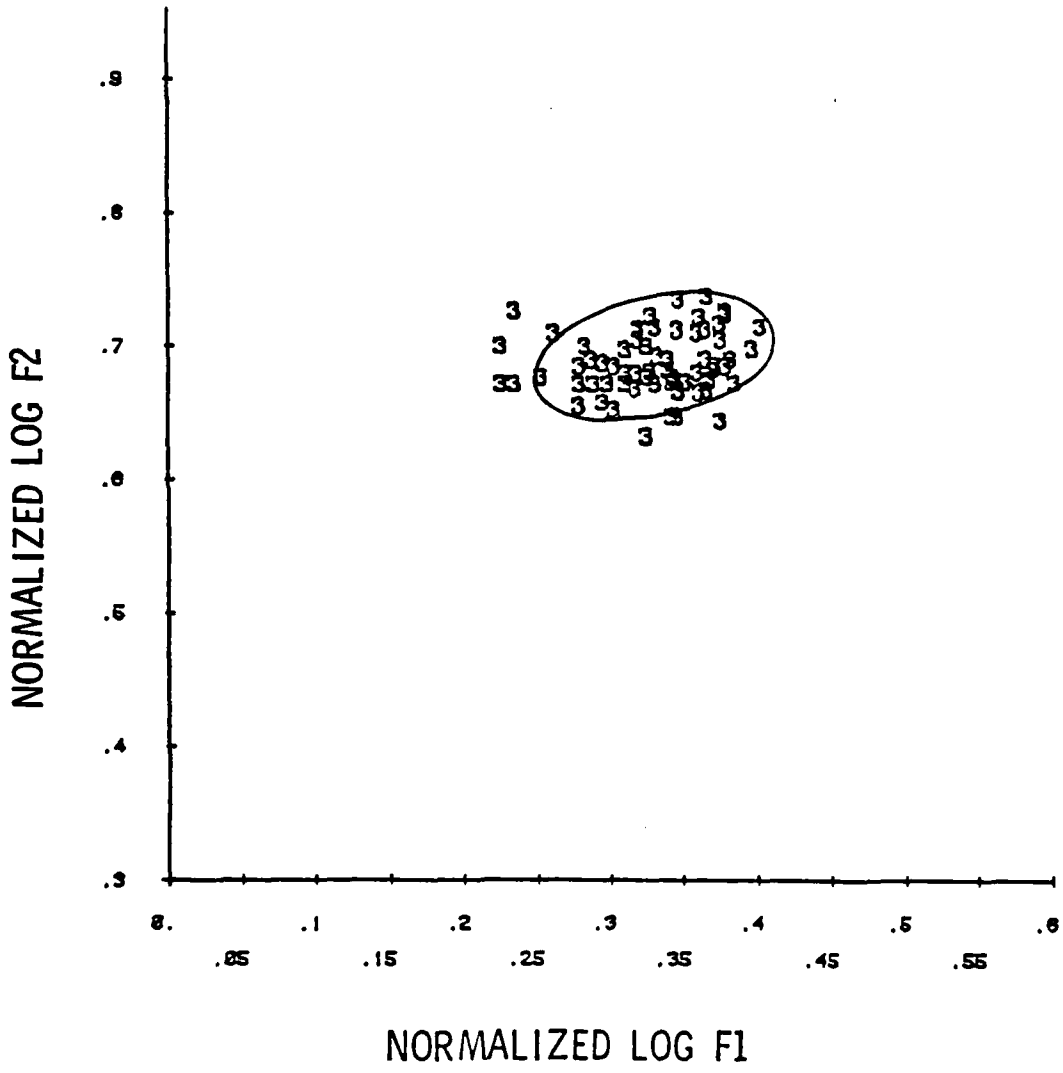


Figure 13. Clustering of the vowel /3/ in the F1-F2 plane.

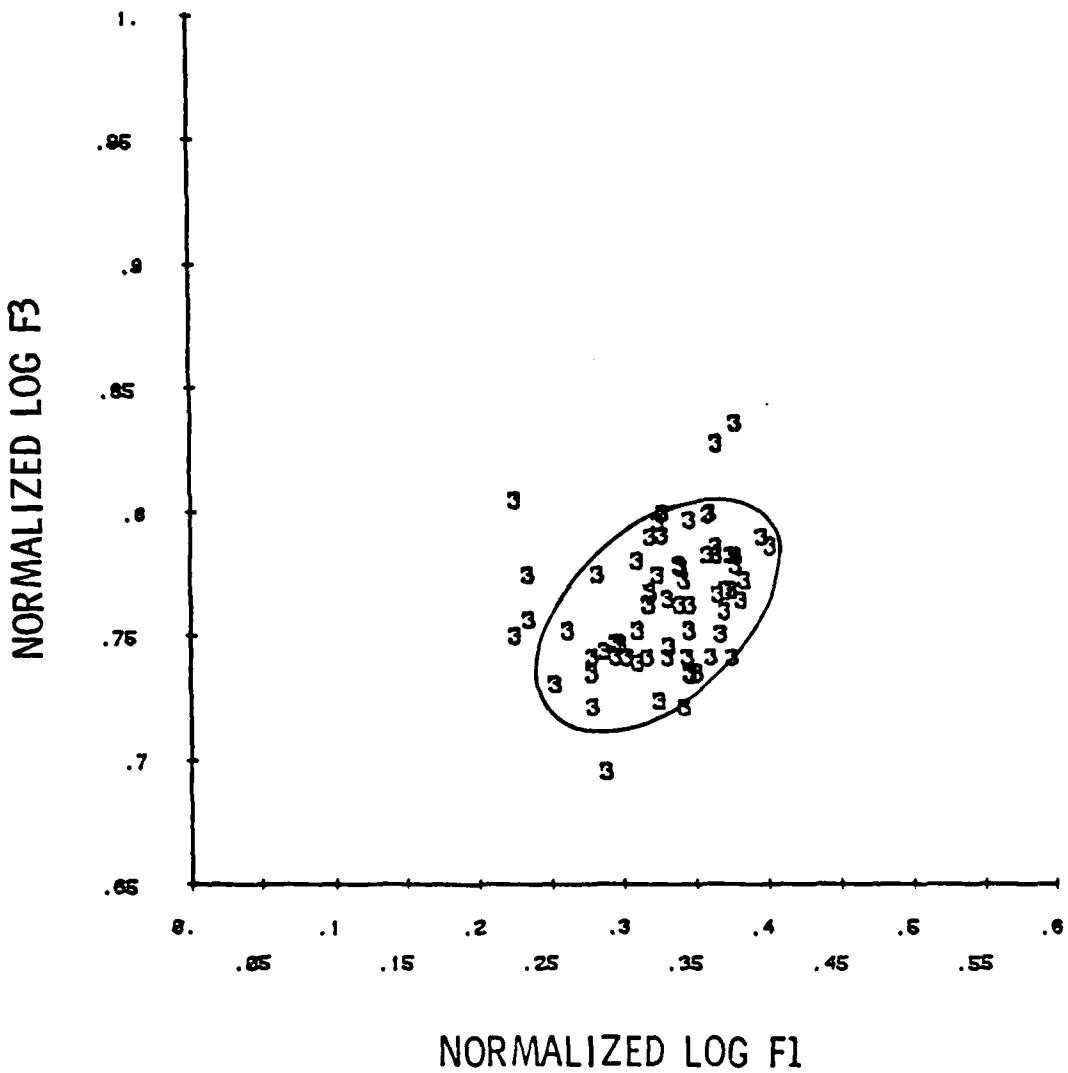


Figure 14. Clustering of the vowel /3/ in the F1-F3 plane.

symbol for the particular vowel which it represents. The symbols are somewhat modified versions of those of the International Phonetic Alphabet, due to the limited availability of IPA symbols for plotting purposes. The results mentioned previously as reported by Peterson, Barney, Potter, and Steinberg are reproduced and verified by the present author. Each vowel cluster on the formant plots is enclosed by a solid line. The exact outline of each cluster is arbitrary; the outlines are intended to indicate a general cluster shape for the purpose of evaluating separability in a graphical, qualitative manner.

Distance measures

Selected quantitative distance measures and cluster sizes are computed for the formant frequencies in two (F1,F2) and three (F1,F2,F3) dimensions (see Appendix A). Although the actual distance measures (Tou & Gonzales, 1974, p. 77) used are arbitrary, a set of dimensionless measurements is necessary to allow comparison of the cluster sizes and vowel separability in the formant space with that in the PARCOR space. The average intracluster distance for each cluster is computed as the average Euclidean distance between each of the data points in the cluster (normalized log frequency) and the centroid of the cluster. The intercluster distance is computed as the Euclidean distance between the centroids of selected adjacent vowel cluster

pairs. Ideally, the average intracluster distances should be minimized for maximum cluster compactness and the intercluster distances should be maximized for maximum cluster separability. The average intracluster distances for each cluster and the intercluster distances for selected adjacent vowel pairs are tabulated in Tables 3 and 4, respectively. The average intracluster distance, a measure of cluster compactness, for each vowel is minimum in two (F1,F2) dimensions, whereas the intercluster distance, a measure of vowel separation, is maximum in three (F1,F2,F3) dimensions. The ratio of the sum of average intracluster distances to intercluster distance for each of the adjacent-vowel pairs is also computed in two and three dimensions, (presented in Table 5). This measurement is meaningful when compared with values computed for the PARCOR coefficients (Chapter 5). The distance measures substantiate the results of the graphical analysis: sufficiency of the F1-F2 plot to identify the first nine vowels and the F1-F3 plot to separate them from the tenth vowel, /3/.

Table 3

Average Intracluster Distances for
Formant Frequency and PARCOR Coefficient Clusters

	Cluster dimension				
Vowel	F(1,2)	F(1,2,3)	K(1,2)	K(1,2,3)	K(1-6)
/i/	0.0456	0.0514	0.1303	0.1418	0.1623
/I/	0.0424	0.0464	0.0848	0.0926	0.1069
/ɛ/	0.0337	0.0375	0.0627	0.0715	0.0834
/æ/	0.0312	0.0360	0.0583	0.0699	0.0830
/ɑ/	0.0263	0.0401	0.0155	0.0650	0.1027
/ɔ/	0.0496	0.0577	0.0116	0.0408	0.0920
/U/	0.0461	0.0533	0.0235	0.0393	0.0737
/u/	0.0775	0.0815	0.0299	0.0463	0.0797
/ʌ/	0.0365	0.0443	0.0264	0.0527	0.0774
/ɜ/	0.0426	0.0504	0.0448	0.0373	0.0567

Table 4

Intercluster Distances for Adjacent-Vowel Pairs of Formant Frequency and PARCOR Coefficient Clusters

	Cluster Dimension				
Vowel pair	F(1,2)	F(1,2,3)	K(1,2)	K(1,2,3)	K(1-6)
i-I	0.1408	0.1482	0.1828	0.1832	0.2510
I-ɛ	0.1143	0.1147	0.0809	0.0917	0.1083
ɛ-æ	0.0852	0.0861	0.0968	0.1067	0.1142
æ-ɝ	0.1388	0.1824	0.1288	0.1842	0.2921
ɑ-ɔ	0.1616	0.1619	0.0676	0.1207	0.1372
ɑ-ʌ	0.0648	0.0653	0.0428	0.0577	0.0625
ɑ-ɝ	0.1598	0.2021	0.0881	0.2291	0.3407
ɔ-U	0.1179	0.1200	0.0539	0.0572	0.1071
ɔ-ʌ	0.1537	0.1537	0.0884	0.1075	0.1364
U-u	0.1376	0.1377	0.0229	0.0542	0.0774
U-ɝ	0.1060	0.1400	0.1155	0.1479	0.2502
ʌ-ɝ	0.0951	0.1495	0.0590	0.1826	0.2936

Table 5

Ratio of the Sum of Average Intracluster Distances to Intercluster Distance for Adjacent-Vowel Pairs of Formant Frequency and PARCOR Coefficient Clusters

	Cluster Dimension				
Vowel pair	F(1,2)	F(1,2,3)	K(1,2)	K(1,2,3)	K(1-6)
i-I	0.6249	0.6599	1.176	1.279	1.072
I-ɛ	0.6661	0.7322	1.824	1.789	1.757
ɛ-æ	0.7624	0.8542	1.251	1.326	1.457
æ-ɝ	0.5312	0.4739	0.7431	0.6223	0.4783
ɑ-ɔ	0.4694	0.6038	0.4005	0.8761	1.419
ɑ-ʌ	0.9695	1.293	0.9804	2.040	2.879
ɑ-ɝ	0.4309	0.4479	0.5996	0.4788	0.4679
ɔ-U	0.8118	0.9247	0.6503	1.401	1.546
ɔ-ʌ	0.5600	0.6637	0.4300	0.8688	1.241
U-u	0.8982	0.9793	1.780	1.579	1.980
U-ɝ	0.8369	0.7408	0.5264	0.5685	0.5209
ʌ-ɝ	0.8311	0.6338	1.081	0.5336	0.4565

CHAPTER III

GENERATION OF SYNTHESIZED VOWEL-LIKE SOUNDS

It was desired to reproduce the data measured at Bell Laboratories as accurately as possible. The speech spectrum may be adequately represented by frequency information below 4000 Hz. (Denes & Pinson, 1963, p. 140) which includes the range of the first three formant frequencies for male speakers. Since each formant must be represented by a complex conjugate pole pair, a six-pole filter is a sufficient representation of a vowel. A sampling frequency of 8000 Hz. is chosen, following Rabiner and Schafer (1978, chap. 3).

Digital Models for the Vocal Tract

Rabiner and Schafer (1978, chap. 3) represent the vocal tract as a recursive IIR digital filter with transfer function

$$H(z) = \frac{1}{1 - \sum_{j=1}^{p} a_j z^{-j}} = \frac{1}{A(z)} = \frac{Z(z)}{V(z)} \qquad (1)$$

where p is the order of the filter. For stability, the z-plane poles corresponding to the roots of this equation must lie inside the unit circle. The corresponding output AR speech process, z(n), is described by the equation:

$$z(n)=a_1 z(n-1)+a_2 z(n-2)+ \ . \ . \ . \ +a_p z(n-p)+v(n)$$
$$=\sum_{j=1}^{p} a_j z(n-j)+v(n) \qquad . \qquad (2)$$

Each formant of a vowel is related to a complex conjugate pair of zeros of the polynomial in z^{-1} , A(z):

$$z_j , z_j^* = e^{-\Delta F_j T} \ e^{+j2\pi F_j T}$$

where F_j is the jth formant frequency, ΔF_j is the bandwidth of the jth formant, and T is the sampling period (Rabiner & Schafer, 1978). So that the filter may be realized recursively, the polynomial coefficients are determined by evaluating the denominator of H(z), where

$$H(z)= \frac{1}{\prod_{j=1}^{p} (1-z_j z^{-1})}$$

and equating the denominator to $A(z)=1-\sum_{j=1}^{p} a_j z^{-j}$.

Physical Model of Speech Production

The physical mechanism for speech production consists of the lungs, bronchi, trachea, larynx, pharynx, and nasal and oral cavities. The larynx, which includes the vocal folds, is the principal structure for voiced speech. Complex tones are produced when short duration air pulses produced at the glottis (the space between the vocal folds) excite the supralaryngeal portion of the vocal tract. Alternately, a noise source may be produced by constricting

the vocal tract (i.e., at the vocal folds, lips, tongue, or soft palate), causing the airstream to become turbulent. For unvoiced speech, the noise source is produced without the vibration of the vocal folds. Here, too, the vocal tract acts as a resonant cavity to shape the resultant sound. Figure 15 (from Flanagan, 1972, p. 24) shows the physical system in terms of the possible mechanisms for sound generation and resonation.

The resonant frequencies of the lossless tube model of the vocal tract have very narrow bandwidths. Actually, the vocal tract is not lossless. The cross section of the vocal tract varies continuously over the length, and energy losses occur as a result of result of viscous friction between air and the walls of the tube, heat conduction through the tube walls, and vibration of the tube walls as well as from losses at the glottis (vocal folds) and lips. These losses are frequency-dependent, and their combined effect is to change the positions of the vocal tract resonances and broaden the bandwidths of those resonances. (Rabiner & Schafer, 1978, p. 72).

Modeling of Formant Bandwidths

The bandwidths of the Peterson and Barney data were measured by Bogert (1953) and again by Dunn (1961). Bogert concluded that bandwidths of formants are invariant and

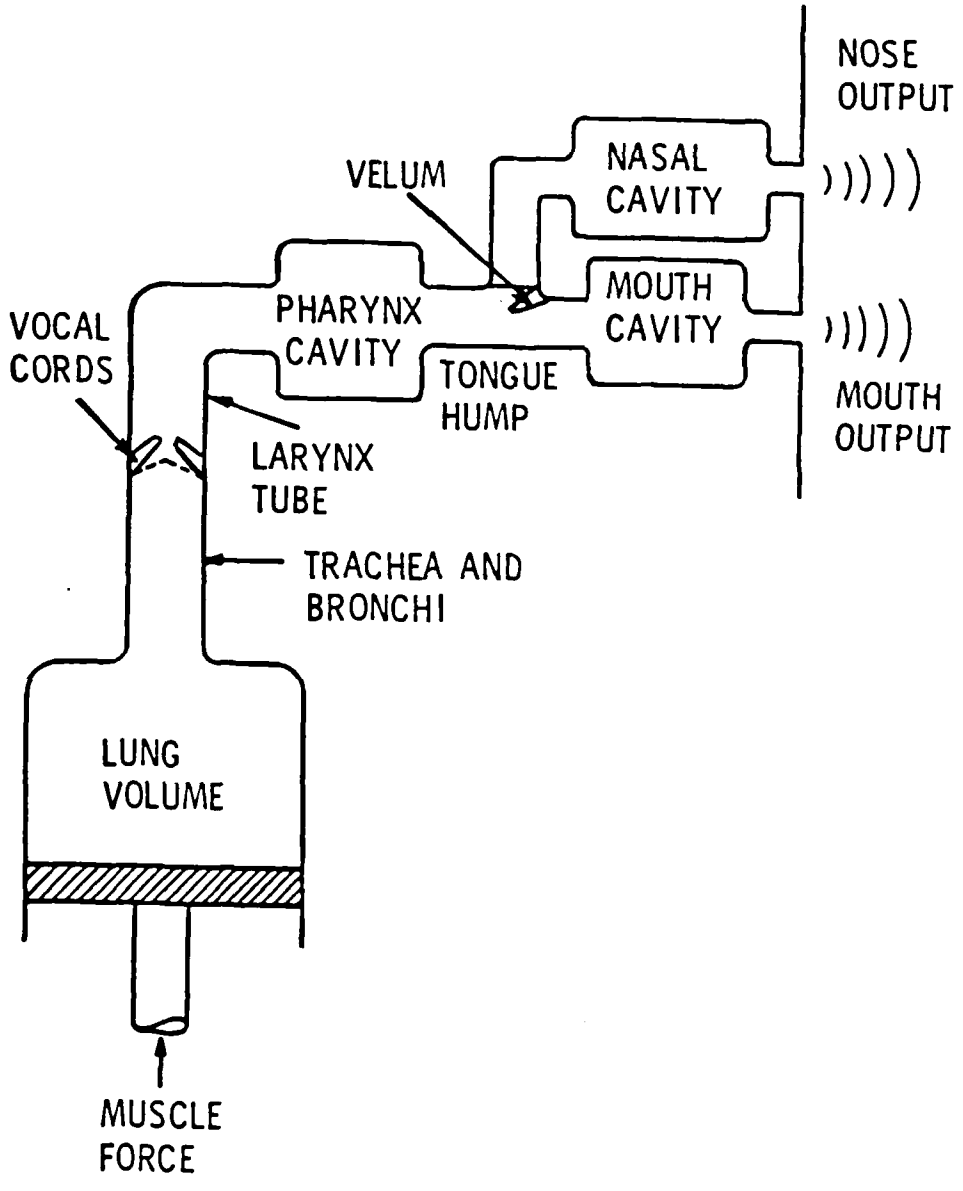


Figure 15. Schematic diagram of functional components of the vocal tract. (From Flanagan, 1972, p. 24)

independent of vowel. Dunn also questioned whether changes in bandwidth from vowel to vowel are critical or even necessary for correct identification of synthetic speech. Neither the bandwidth data measured by Dunn nor that measured by Bogert were available for use in this study. The required bandwidth values for each formant frequency are supplied by averages over all ten vowels ($\Delta F_1 = 52.0$, $\Delta F_2 = 66.0$, $\Delta F_3 = 120.0$) from sine wave bandwidths for synthesized vowels determined by Dunn with his electrical vocal tract.

A series of experiments conducted by this author describes the effects of bandwidth changes on the PARCOR coefficients of a single formant (two-pole) system. For a series of single formant frequency systems with formants incremented over the range 250-3500 Hz., when the bandwidth of the formant is incremented from 10-200 Hz., the range (averaged over the formant frequency experiments) over which K1 varies is .0057 with a standard deviation of .0034. The average range over which K2 varies is .0115 with a standard deviation of .1361. Likewise, for a series of constant-bandwidth single-formant frequencies, when the formants are incremented over the range 250-3500 Hz. for each bandwidth in the range 10-200 Hz. The range (averaged over the bandwidth experiments) over which K1 varies is 1.900 with a standard deviation of .0040; the average range of K2 is

.0061 with a standard deviation of .0255. In comparing the ranges of K1 and K2, it is observed that the effect of formant frequency variations on the PARCOR coefficients, especially K1 (which has been previously determined by Tohkura & Itakura, 1979, to be sensitive to variations in pole placement) is much more pronounced than the effect of bandwidth variations, which are practically negligible. Incidentally, it is noted that as ΔF_j is increased for any F_j, K2 decreases. As F_j is increased for any ΔF_j, K1 increases.

Modeling of Vocal Tract Excitation

Modeling of the vocal tract excitation function presents a problem with respect to the generation of the synthesized vowel-like sounds. The primary objective in reproducing the Bell Laboratories data is to accurately reproduce the formant frequencies measured by Peterson and Barney. Input to the vowel generation prefilter must be white in order to obtain an AR process as the output. Ideally, the easiest way to exactly specify the frequency peaks (formants) of the output speech spectrum is to specify them as the peaks of the transfer function of the filter, using an input signal with a constant (flat) frequency spectrum. This requirement of a flat spectrum is satisfied by two types of signals; white noise and a periodic (deterministic) impulse train (corresponding to unvoiced and

voiced excitation, respectively). Although the frequency spectrum of an impulse train is flat, modeling of the excitation as an impulse train is less than desirable in terms of the accuracy with which the harmonics of the impulse frequency correspond to the desired frequency peaks (formants) of the output signal.

The speech spectrum for a vowel should theoretically have formant frequencies which are integer multiples of the fundamental frequency. However, the fundamental and formant frequencies measured by Peterson and Barney do not exhibit this tendency for several reasons. Primarily, the technique used by Peterson and Barney to measure formant frequency used a weighted average of the frequency components of the spectral peak (Potter & Steinberg, 1950); in addition, factors such as perturbation in actual fundamental frequency and difficulty in interpreting spectrograms to within a few Hertz, as well as measurement and roundoff error probably contributed to the lack of relationship between measured fundamentals and their formants. Since the measured fundamentals are fairly low in frequency, their harmonics are sufficiently far apart as to cause the reproduced formants to be much different from the measured (desired) formants.

As stated earlier, the primary objective in the generation of the data is to recreate the measured formants.

Therefore, the fundamental frequency data are disregarded, and the input excitation is chosen to be a white noise sequence with zero mean and standard deviation equal to one. This may be likened to an unvoiced input, as though whispered vowels are being generated. Although the power spectral density of the excitation function of a whispered vowel is not exactly white, the approximation of the input as white noise is no less accurate from a signal processing standpoint than approximating the glottal wave (voiced input) with an impulse train.

The input excitation problem is illustrated in Figure 16. A voiced input is shown in Figure 16(b) as a periodic, deterministic impulse train of frequency f_0 with harmonics f_n, where $f_n = nf_0$. An unvoiced input is modeled in Figure 16(d) as white noise, with frequency components at f_n, (limited by the sampling frequency F_s), where $f_n = n/F_s$. Comparing the output spectra of Figure 16(c) and Figure 16(e), the reproduced formants $\tilde{F}_n$ are closer to the desired formants F_n, for the noise input of Figure 16(d) than for the impulse input of Figure 16(b).

Each synthesized vowel-like utterance is generated as a sixth-order AR time series from the three formant frequencies supplied by the Bell Laboratories data (Barney, 1952). A Gaussian white noise process, v(n), with $\sigma_v^2 = 1.0$ is used as the excitation function. Bandwidths for F1, F2,

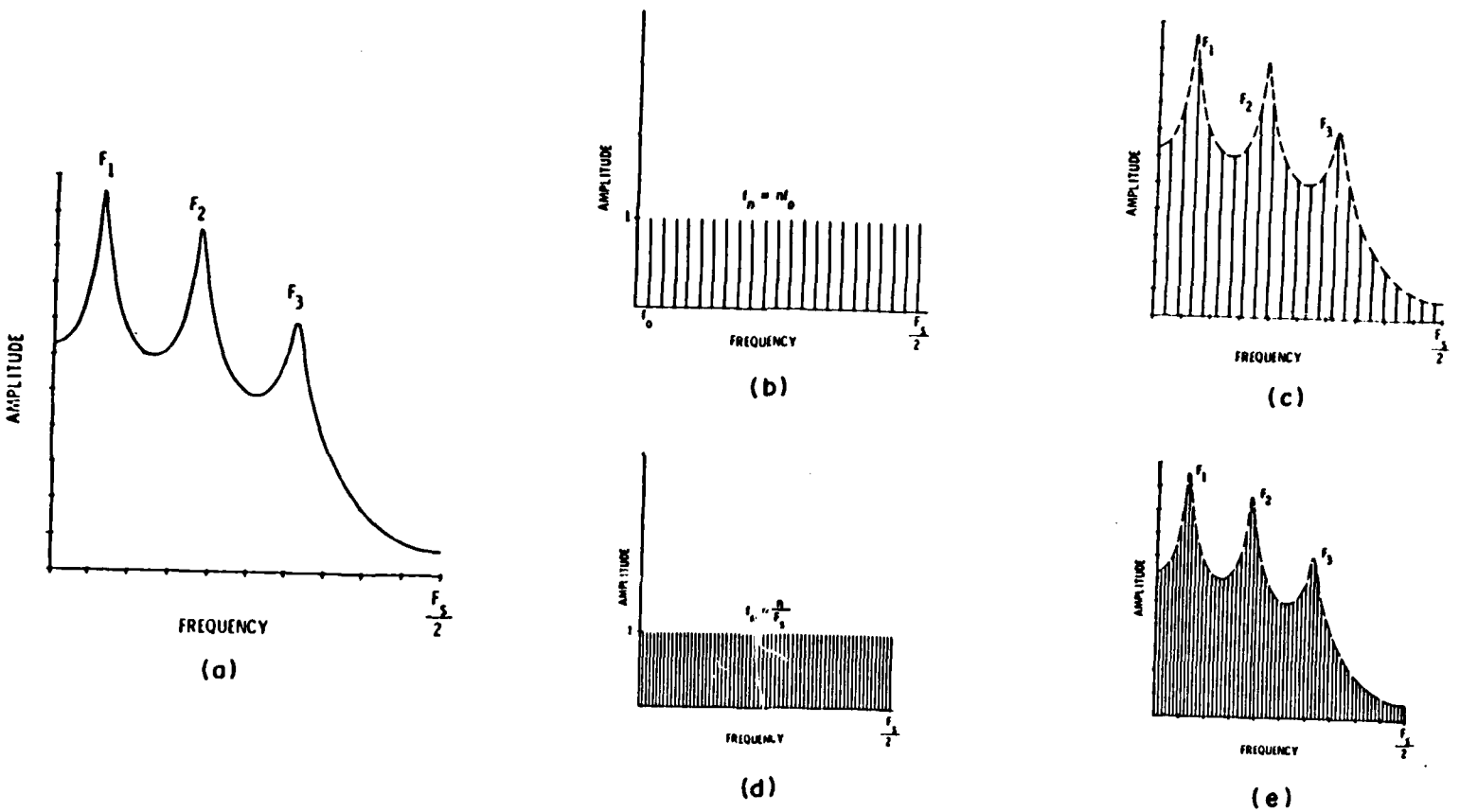


Figure 16. Synthesis of vowel sound via excitation of transfer function with voiced and unvoiced white input. (a) Desired power spectrum of synthesized vowel. (b) Power spectrum of deterministic, periodic, voiced input. (c) Actual power spectrum of synthesized voiced vowel. (d) Power spectral density of unvoiced input. (e) Actual power spectral density of synthesized unvoiced vowel.

and F3 are held constant for all utterances at the values 52.0 Hz., 66.0 Hz., and 120.0 Hz., respectively. The process is generated for 1000 samples. These samples are then fed into the sixth-order inverse whitening filter (see Chapter IV) to obtain the PARCOR coefficients which are used as pattern recognition parameters for vowel identification. A system block diagram of the computer simulation is shown in Figure 17.

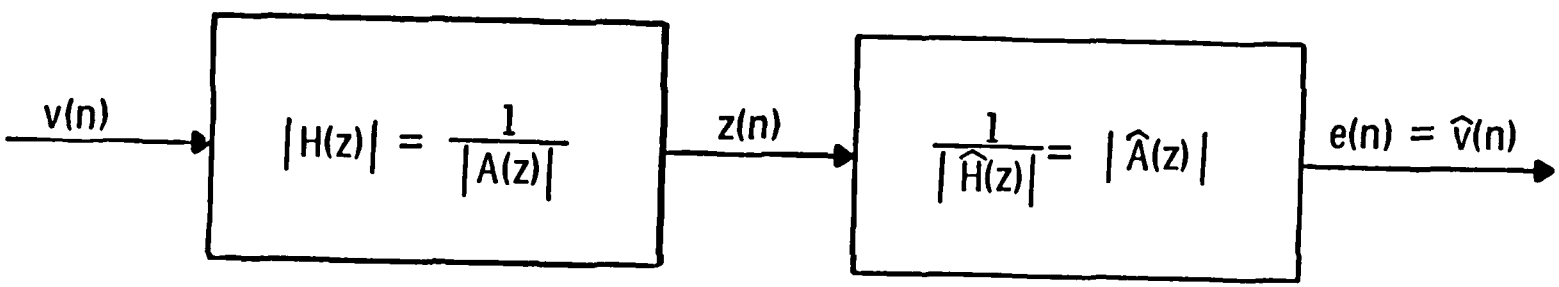


Figure 17. System block diagram for the prefilter-lattice filter sequence.

CHAPTER IV

INVERSE FILTER

The forward PARCOR coefficients $K_i^e(n)$, $i=1, 2, \ldots 6$ corresponding to each utterance are obtained by passing 1000 of the synthesized vowel-like samples, $z(n)$, through a sixth-order inverse (whitening) filter. The six final forward PARCOR coefficients $K_i^e(1000)$ are used in Chapter 5 as pattern recognition parameters for the vowel utterances.

The Linear Prediction Problem

Rabiner and Schafer (1972, chap. 8) present the linear prediction problem; for the digital vocal tract model of equation 1, the speech samples, $z(n)$, are related to the white input samples, $v(n)$, by equation 2. A linear predictor with predictor coefficients, $\hat{a}_j$, is defined as $z(n)= \sum_{j=1}^{p} \hat{a}_j z(n-j)$. The system function for this pth-order linear predictor is $P(z)= \sum_{j=1}^{p} \hat{a}_j z^{-j}$. The prediction error, $e(n)$, is the difference between the speech sample, $z(n)$, and the linearly predicted one, $\hat{z}(n)$:
$e(n)= z(n)-\hat{z}(n)= z(n)- \sum_{j=1}^{p} \hat{a}_j z(n-j)$. The error, $e(n)$, is the output of a system with transfer function

$$\frac{1}{\hat{H}(z)} = \hat{A}(z) = 1- \sum_{j=1}^{p} \hat{a}_j z^{-j} .$$

which recovers the white input ($e(n)=\hat{v}(n)$) by removing the

correlation between samples of z(n). The error, e(n), will approach v(n) as $\hat{a}_j$ approach a_j and $\hat{A}(z)$ will be the inverse filter for H(z). The linear prediction problem then is one of finding the $\hat{A}(z)$ which minimizes the square of the exponentially weighted forward prediction error given by $E_p(N)= \sum_{n=0}^{N} (1-\alpha_{CLSL})^{N-n} |e_p(n)|^2$ (Hodgkiss & Presley, 1981). This leads to a set of linear equations called the normal equations. Complete algebraic derivations of the least squares lattice equations (from which this was adapted) are found in Lee (1980) and Pack and Satorius (Note 4).

The Least Squares Lattice

The solution of the normal equations is dependent on the efficient inversion of a pxp covariance matrix (Lee, 1980). Several solution algorithms are discussed by Morf, Lee, Nickolls, and Vieira (1977). The Levinson (1947) algorithm is an efficient least squares simultaneous solution to the normal equations requiring only $O(p^2)$ computations per time update (where p is the order of the filter) for a stationary process. A natural implementation of Levinson's algorithm, the lattice structure (as realized by Itakura and Saito, 1971), provides an extension to the non-stationary case. Lee (1980) and Pack and Satorius (Note 4) present Levinson's recursion clearly. A class of fast exact least squares algorithms which require only O(p) computations per time update are discussed in the literature

by Morf, Dickenson, Kailath, and Vieira (1977), and Morf, Lee, Nickolls, and Vieira (1977). An exact time update recursion in terms of lattice variables only has been developed and tested by Lee (1980), Morf and Lee (1978), Morf, Lee, Nickolls, and Vieira (1977), and Morf, Vieira, and Lee, (1977).

Lattice Structure

A feed forward (MA) lattice structure (Figure 18) is the realization of Levinson's algorithm for the computation of the optimal linear predictor. The lattice is composed of a cascade of p individual lattice sections, corresponding to the stages (order) of the algorithm, $i=1,2, \ldots p$. The variable in the lower path, $e_i(n)$, is the forward error between the input, $z(n)$, and the least squares (linearly predicted) estimate of $z(n)$, $\hat{z}(n)$, based on a linear combination of past inputs: $\hat{z}(n)=\sum_{j=1}^{p} \hat{a}_j z(n-j)$. Likewise, the backward prediction error, $r_i(n)$, propagates backward along the upper path. The variables represented by the cross bars of the lattice are the forward and backward partial correlation or PARCOR coefficients which arise naturally as intermediate entities in the solution of the Levinson algorithm. For the least squares lattice, $K_i^e \neq K_i^r$ and $|K_i^e, K_i^r| < 1$ for $i=1,2, \ldots p$.

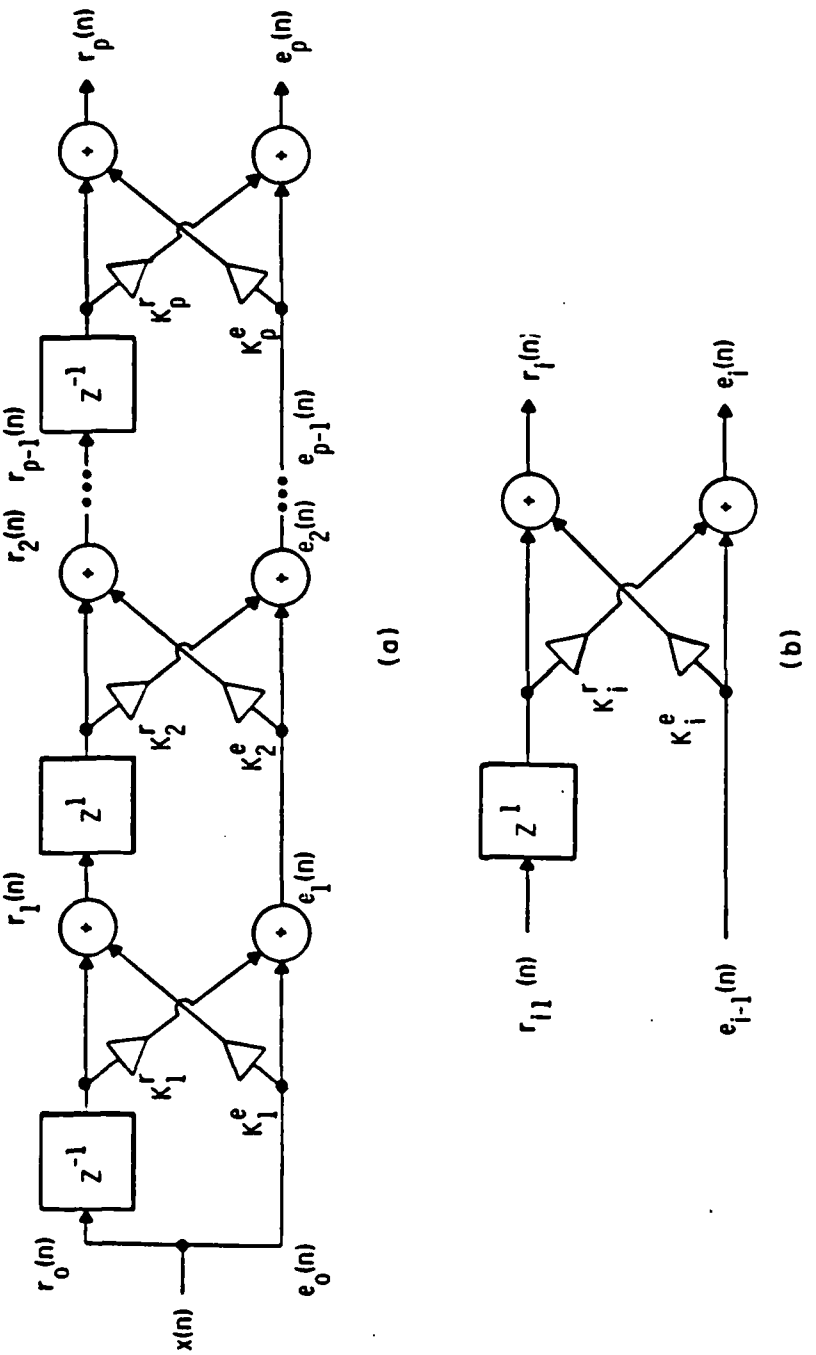


Figure 18. (a) Forward and backward prediction error filters. (b) The ith stage of the lattice. (From Hodgkiss & Presley, 1982, p. 331)

Lattice Variables

Appendix B presents the least squares complex adaptive lattice variables and equations as adapted from Lee, Morf, and Friedlander (1981) for complex data by Hodgkiss and Presley (1982). The software supplied by Alexandrou and Hodgkiss (Note 1) for the computer simulation is based on these equations.

Fade factor. The fade factor, $(1-\alpha_{CLSL})$, applies an exponential weighting on the data by weighting recent errors more heavily than those in the distant past. For this study the value of α_{CLSL} was chosen to be .0001, although the choice is not critical here as the time series are all stationary. Bounded by [0,1], $(1-\alpha_{CLSL})$ is usually close to 1; the inverse of α_{CLSL} is approximately the memory of the algorithm (Pack & Satorius, Note 4). The value of α_{CLSL} may be selected to satisfy a misadjustment criterion (Hodgkiss & Presley, 1982).

Likelihood variable. A major difference between gradient lattice developed by Griffiths (1975, 1977) and LSL algorithms is the Gaussian likelihood parameter, $\gamma_{i-2}(n-1)$, which replaces the constant step size of the gradient lattice and is responsible for the fast tracking capabilities of the LSL. For likely samples (the lower bound of $\gamma_{i-2}(n-1)=0$ is reached for $\alpha_{CLSL}=0$), the step size

is small and constant, roughly on the order of magnitude of the "optimal" gradient step size. For unlikely samples (the upper bound of $\gamma_{i-2}(n-1)=1$ is reached for $\alpha_{CLSL}=1$), $\gamma_{i-2}(n-1)$ will approach unity; the gain, $1/(1-\gamma_{i-2}(n-1))$, is very large, causing the lattice parameters to change quickly to adapt to sudden changes in the input data (Hodgkiss & Presley, 1981). The values of both α_{CLSL} and $\gamma_{i-2}(n-1)$ will become critical if the study is extended such that we proceed to examine the time-varying behavior of the PARCOR coefficients for a nonstationary input time series.

Partial correlation (PARCOR) coefficients. The variable $\Delta_i(n)$ is known as the ith-order partial autocorrelation between z(n) and z(n-i-1), and is defined as the correlation betweeen these two samples after removing their mutual linear dependence on intervening samples. The partial correlation (PARCOR) coefficients K_i^e and K_i^r are the partial autocorrelations normalized by $E_{i-1}^e(n)$ and $E_{i-1}^r(n-1)$.

Performance Measures for the Lattice

Various performance measures may be employed to evaluate the whitening properties of the lattice and the accuracy with which the predictor coefficients, $\hat{a}_j$, identify the transfer function of the prefilter, assuming a white input to the prefilter. Two example cases are presented in

Table 6, for the vowels /i/ and /u/. The pth-stage mean square error (Hodgkiss & Presley, 1982), $E[|e_p(n)|^2]$, may be plotted at each time n. As a quantitative measure of convergence time, the 10% settling time may be determined as the time at which the mean square error comes within 10% of the theoretical steady state value. The mean square error for Example 1 is plotted in Figure 19. The final (pth-stage) error power (Hodgkiss & Presley, 1982) after convergence, $E[|e_p(n)|^2]=E[|\hat{v}(n)|^2]=\sigma_e^2$, will be an estimate of the variance of the prefilter input signal, $E[|v(n)|^2]=\sigma_v^2$. For zero-mean Gaussian white noise input with $\sigma_v=1$, σ_e^2 should approach one for a true whitening filter. The misadjustment, $|\sigma_e^2-\sigma_v^2/\sigma_v^2|$, is also a popular performance measure. The misadjustment after 1000 samples is 8% and 9% for Example 1 and Example 2, respectively.

When the filter transfer functions are realized from the filter coefficients, the plot of the transfer function is a good qualitative performance measure. Inverted, the magnitude of the transfer function of the lattice filter, $|\hat{A}(z)|$, is an approximation of the prefilter (vocal tract model) transfer function, $|1/A(z)|$. This series of transfer functions is presented in Figure 20 for Example 1 and in Figure 21 for Example 2.

The power spectral density may also be used as another qualitative performance measure. For a prefilter input

Table 6
Selected Filter Parameters for
Two Example Vowel Utterances

	j					
	1	2	3	4	5	6
	Example 1, vowel /i/					
F	244	2300	2780	---	---	---
F	52	66	120	---	---	---
a	-0.3717	0.3330	-1.592	0.3281	-0.2648	0.8295
a (N)	-0.3604	0.3449	-1.585	0.3332	-0.2635	0.8145
K (N)	-0.2498	0.0314	-0.8441	0.1920	0.0860	0.8125
Kj	0.3653	0.5084	0.0628	0.5902	0.5362	0.9060
Final mean square error $E[\|e_6(1000)\|^2] = 0.9178 = .3723$ dB						
10% settling time = 757						
	Example 2, vowel /u/					
F	340	950	2240	---	---	---
F	52	66	120	---	---	---
a	-2.964	4.337	-4.541	3.978	-2.558	0.8295
a (N)	-2.919	4.224	-4.381	3.817	-2.459	0.8077
K (N)	-0.9255	0.8691	-0.6304	0.4167	-0.2982	0.8006
Kj	0.0214	0.9349	0.1716	0.7046	0.3407	0.9000
Final mean square error $E[\|e_6(1000)\|^2] = .9119 = .4003$ dB						
10% settling time = 766						

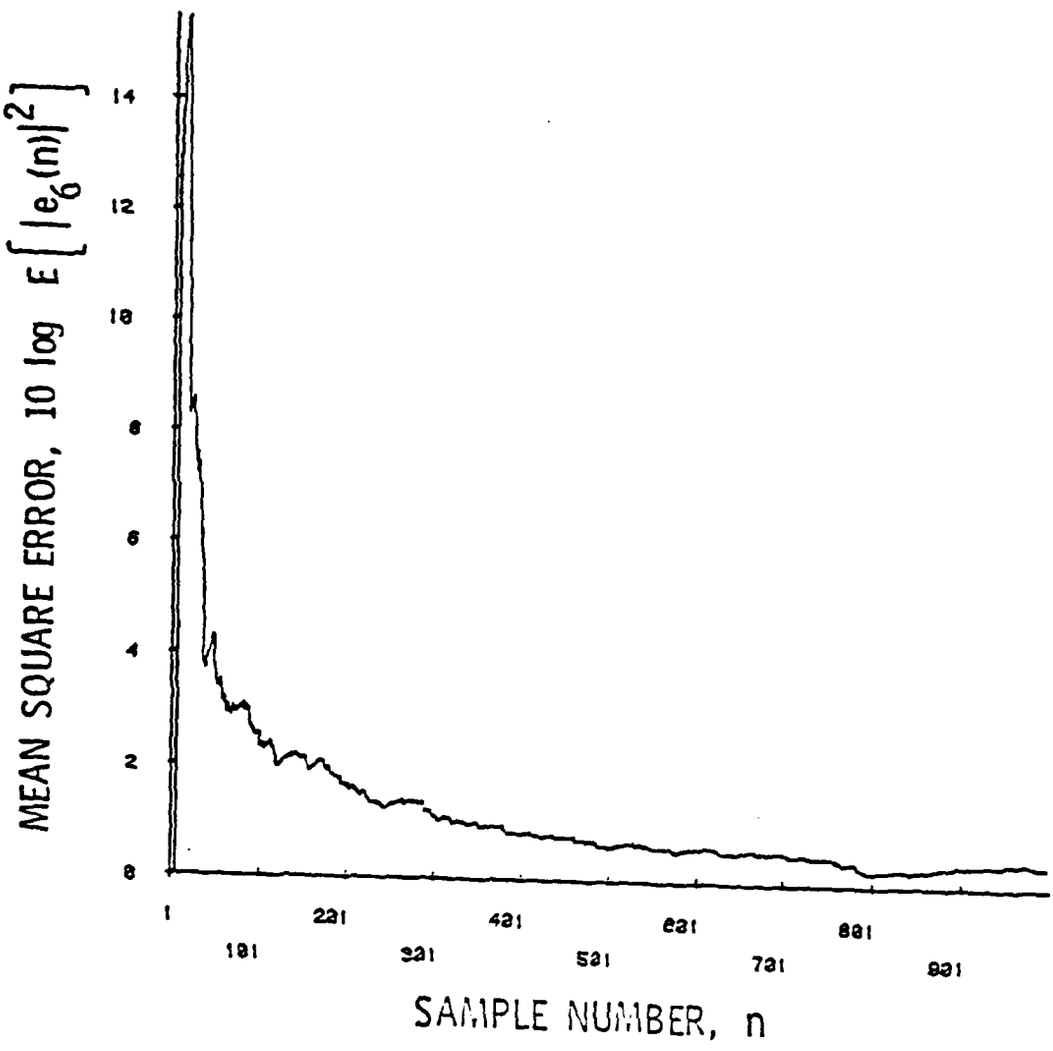


Figure 19. Mean square error, 10 log $E[|e_6(n)|^2]$, for Example 1 in Table 6.

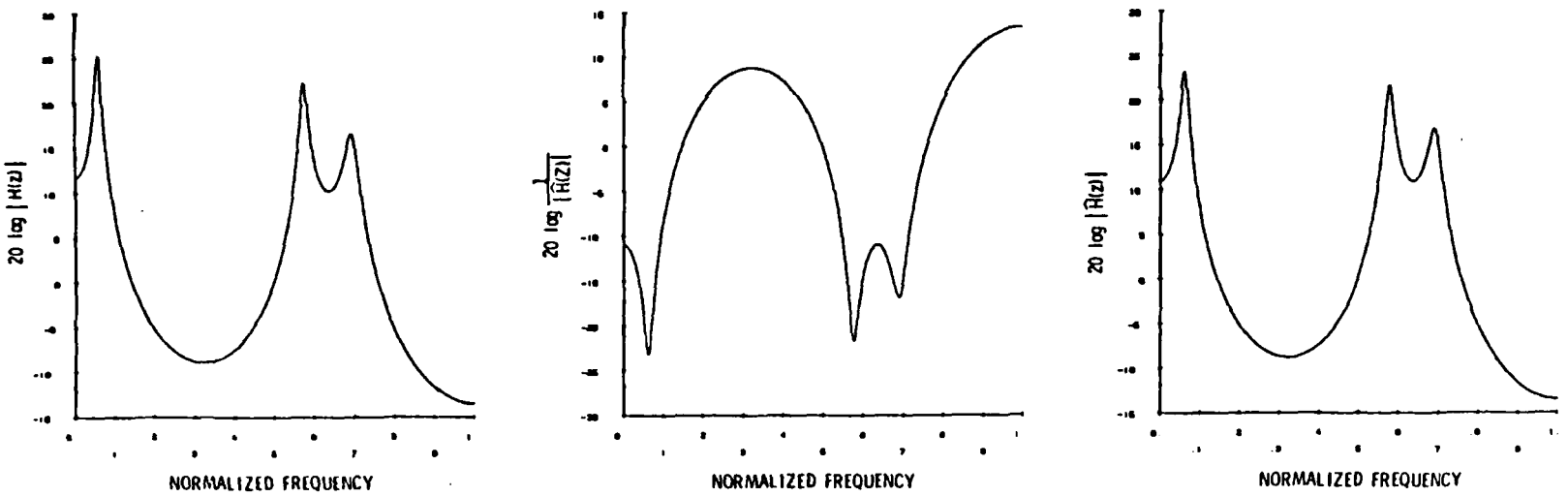


Figure 20. Transfer functions for Example 1, Table 6. (a) Prefilter transfer function, $|H(z)|$. (b) Lattice filter transfer function, $1/|\hat{H}(z)|$. (c) Inverted lattice filter transfer function, $|\hat{H}(z)|$, is an estimate of the prefilter transfer function H(z).

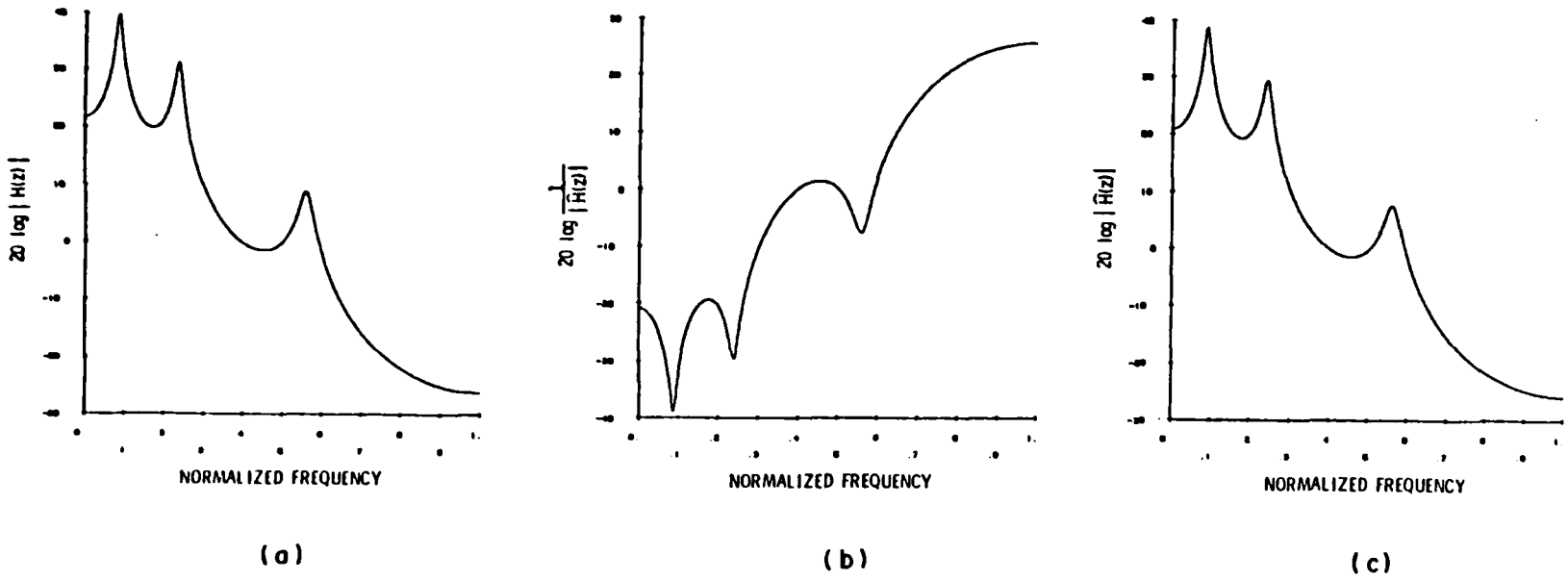


Figure 21. Transfer functions for Example 2, Table 6. (a) Prefilter transfer function, $|H(z)|$. (b) Lattice filter transfer function, $1/|\hat{H}(z)|$. (c) Inverted lattice filter transfer function, $|\hat{H}(z)|$, is an estimate of the prefilter transfer function H(z).

which is zero-mean Gaussian white noise, the power spectral density of the prefilter input is proportional to σ_v^2, a constant: $P_v(z) = 10 \log C\sigma_v^2$. The prefilter output (lattice input) power spectral density is given by Papoulis (1981) as

$$P_z(z) = 10 \log \frac{C\sigma_v^2}{|A(z)|^2} .$$

An estimate of the prefilter input power spectral density is the power spectral density of the lattice output $P_e(z) = 10 \log P_z(z)|\hat{A}(z)|^2$. It should be flat (constant) for optimal whitening. The power spectral density series for the two examples are shown in Figures 22 and 23. An estimate of the prefilter output power spectral density is given by Griffiths (1975) as

$$\hat{P}_z(z) = 10 \log \frac{C\sigma_e^2}{|\hat{A}(z)|^2} .$$

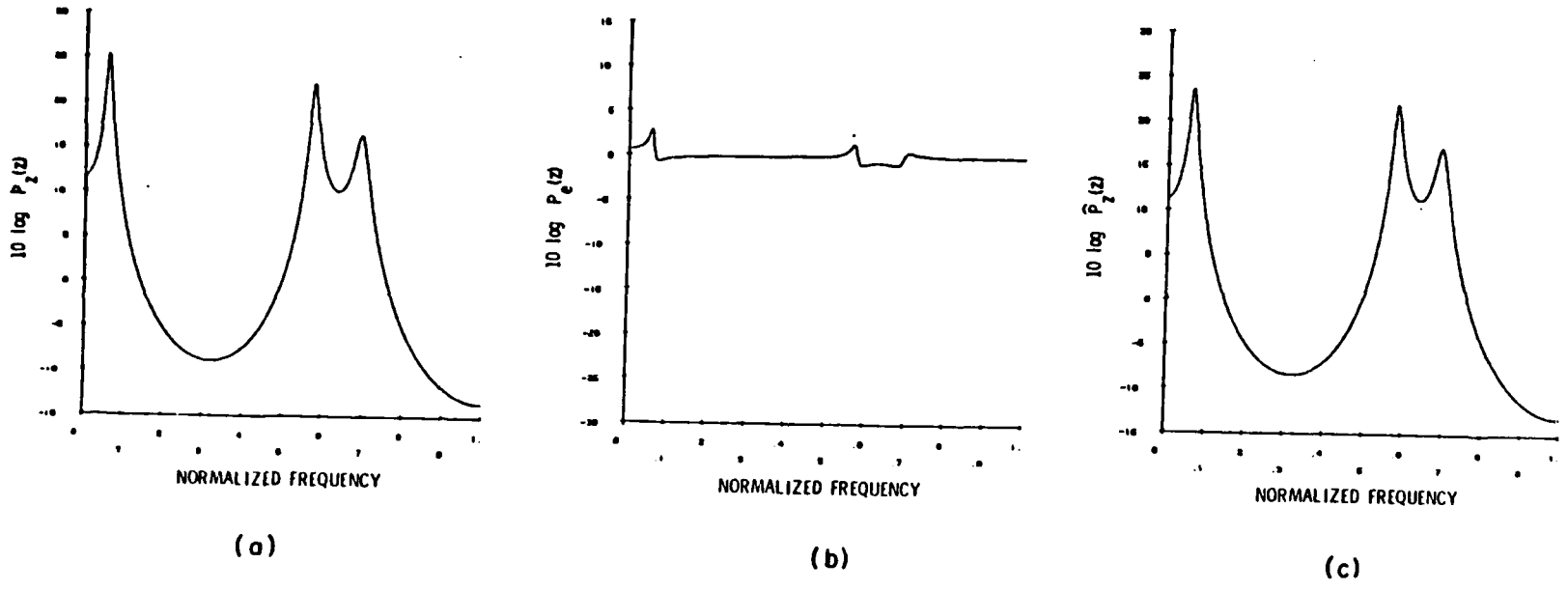


Figure 21. Power spectral densities for Example 1, Table 6. (a) Prefilter output power spectral density $P_z(z)=C\sigma_v^2/|A(z)|^2$. (b)Lattice filter output power spectral density $P_e(z)=P_z(z)|\hat{A}(z)|^2$, is an estimate of the white prefilter input power spectral density $P_v(z)=C\sigma_v^2$. (c) An estimate of the prefilter output power spectral density, $P_z(z)$, is $\hat{P}_z(z)=C\sigma_e^2/|\hat{A}(z)|^2$.

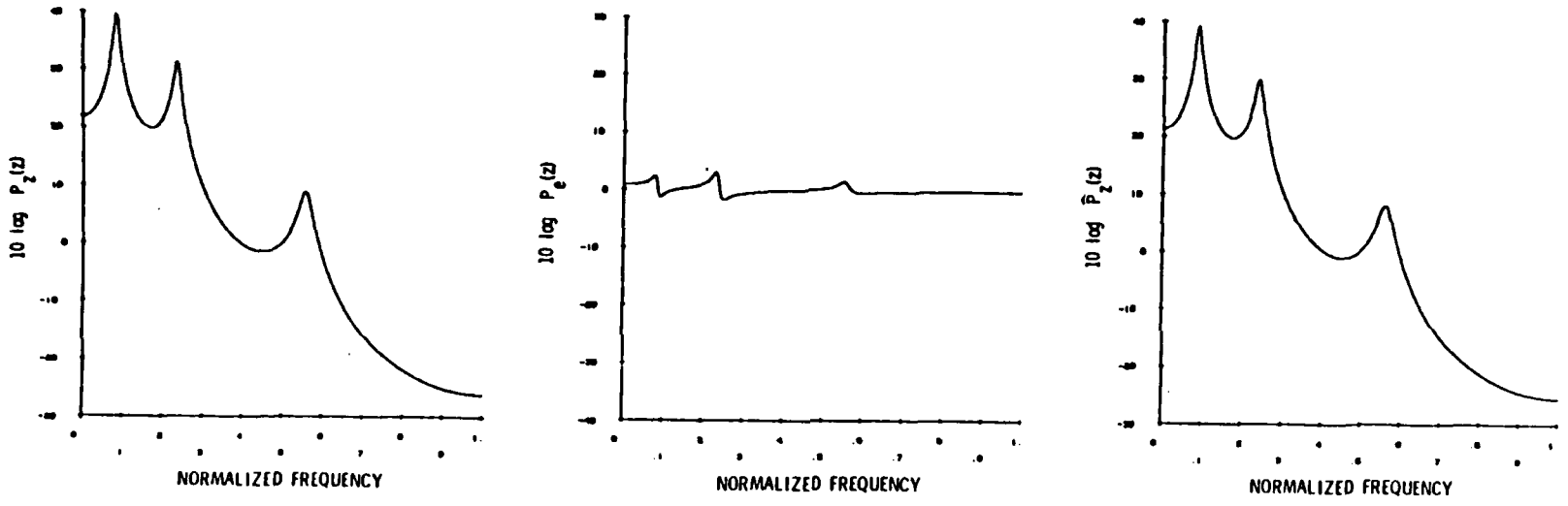


Figure 23. Power spectral densities for Example 2, Table 6. (a) Prefilter output power spectral density $P_z(z)=C\sigma_v^2/|A(z)|^2$. (b)Lattice filter output power spectral density $P_e(z)=P_z(z)|\hat{A}(z)|^2$, is an estimate of the white prefilter input power spectral density $P_v(z)=C\sigma_v^2$. (c) An estimate of the prefilter output power spectral density, $P_z(z)$, is $\hat{P}_z(z)=C\sigma_e^2/|\hat{A}(z)|^2$

CHAPTER 5

RESULTS

For each synthesized utterance, the least squares lattice computes a set of PARCOR coefficients at each time update. The six forward PARCOR coefficients at the last time update $K_i^e(N)=K_i^e(1000)$ obtained as the output of the lattice, are the classification parameters in this study. Because the PARCOR coefficients span the range [-0.9674, 0.9971], they are normalized to span the interval [0,1] to facillitate comparisons of distance measures and cluster sizes between formant clusters and PARCOR clusters. Henceforth, the notation K1, K2, . . . K6 will denote normalized $K_i^e(1000)$; i=1, 2, . . . 6. Table 2 lists the normalized ranges of all the PARCOR coefficients.

Analysis of the PARCOR Coefficient Data

The PARCOR coefficients for all of the synthesized utterances are analyzed both graphically and quantitatively in the same manner as the formant frequencies were analyzed. Distance measures are computed in two (K1,K2), three (K1,K2,K3), and six (K1-K6) dimensions in a manner analogous to the computation of the two (F1,F2) and three (F1,F2,F3) dimensional formant frequency distance measures. The data for the synthesized vowel-like sounds are shown in the K1-K2, K1-K3, K1-K4, K1-K5, and K1-K6 planes in

Figures 24-28.

Graphical Representation

From a graphical analysis, vowels are most separable in the K1-K2 plane, except for the vowel /ɜ/, for which K3 is quite low relative to that of the other vowels and which may be differentiated by its location in the three-dimensional space defined by K1,K2 and K3. Figures 29-38 present the synthesized vowel-like sounds, singly, in the K1-K2 plane. The vowel /ɜ/ in the K1-K3 plane is shown in Figure 39. Graphically, none of the other PARCOR coefficients further separates the vowels. As for the formant space vowel clusters, the precise vowel cluster areas enclosed on the PARCOR plots are arbitrary, intended to indicate a general cluster shape for the purpose of evaluating separability in a graphical, qualitative manner. The range of K2 is greatest, followed by that of K1 (as for the formant frequencies). K1 includes the highest value of PARCOR coefficient; K2 includes the lowest. The range of K6 is the smallest of the PARCOR ranges. The ranges spanned by the various PARCOR coefficients are in accordance with the results of Tohkura and Itakura (1979) who noted that the spectral sensitivity for the first PARCOR is often quite high, and its distribution is wider than that of the higher order PARCOR coefficients.

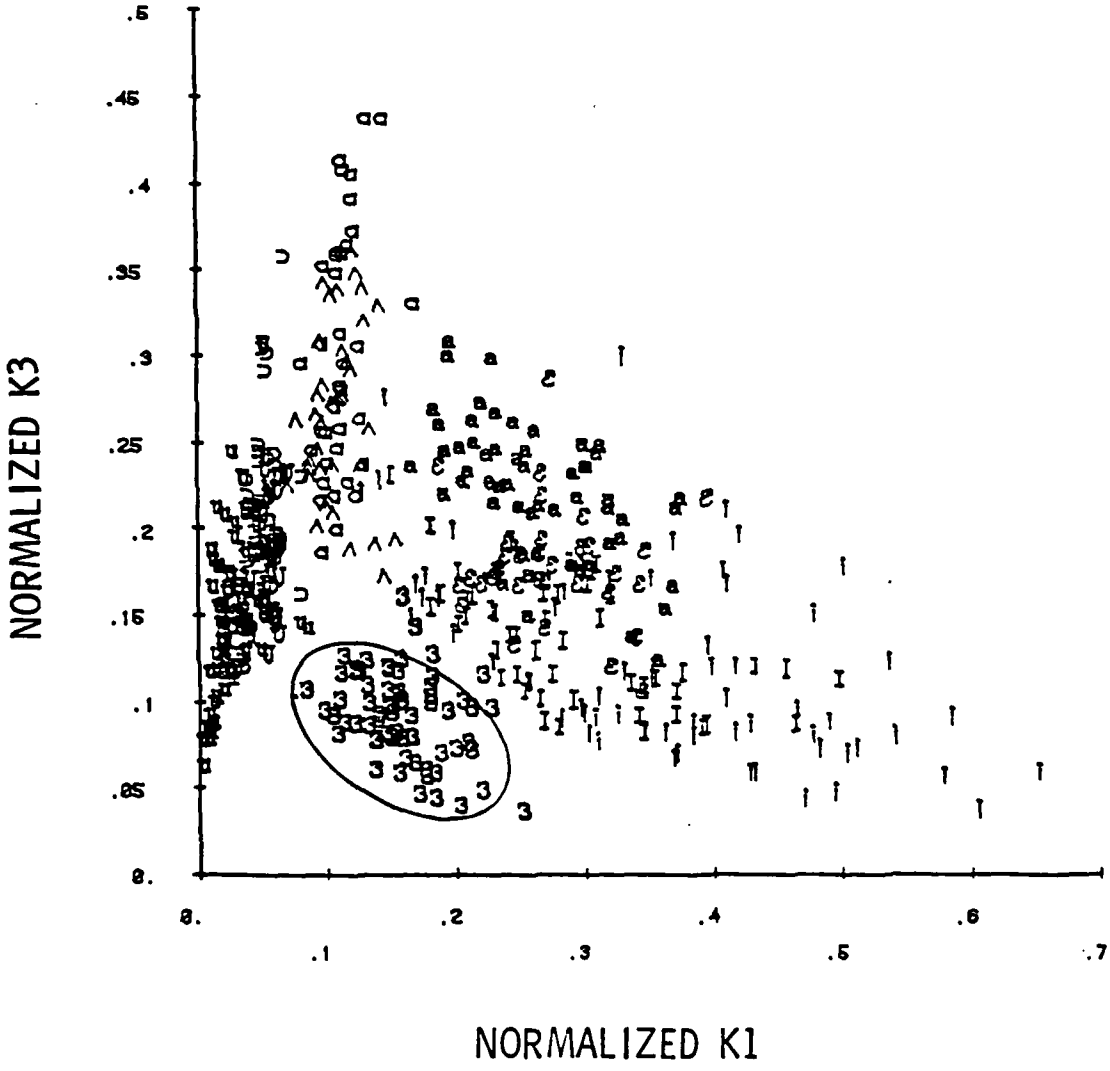


Figure 24. Clustering of the ten English vowels in the K1-K2 plane.

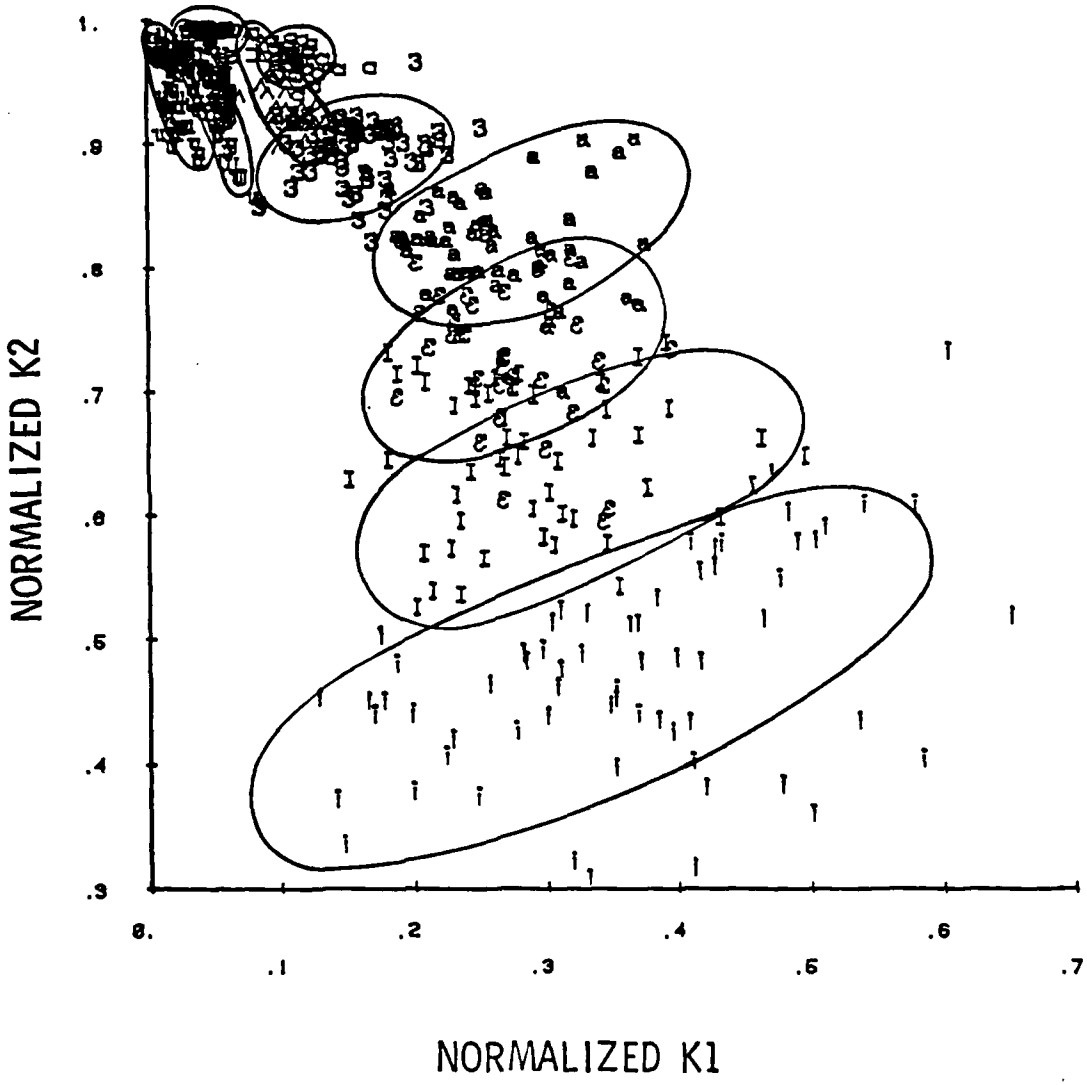


Figure 25. Clustering of the ten English vowels in the K1-K3 plane.

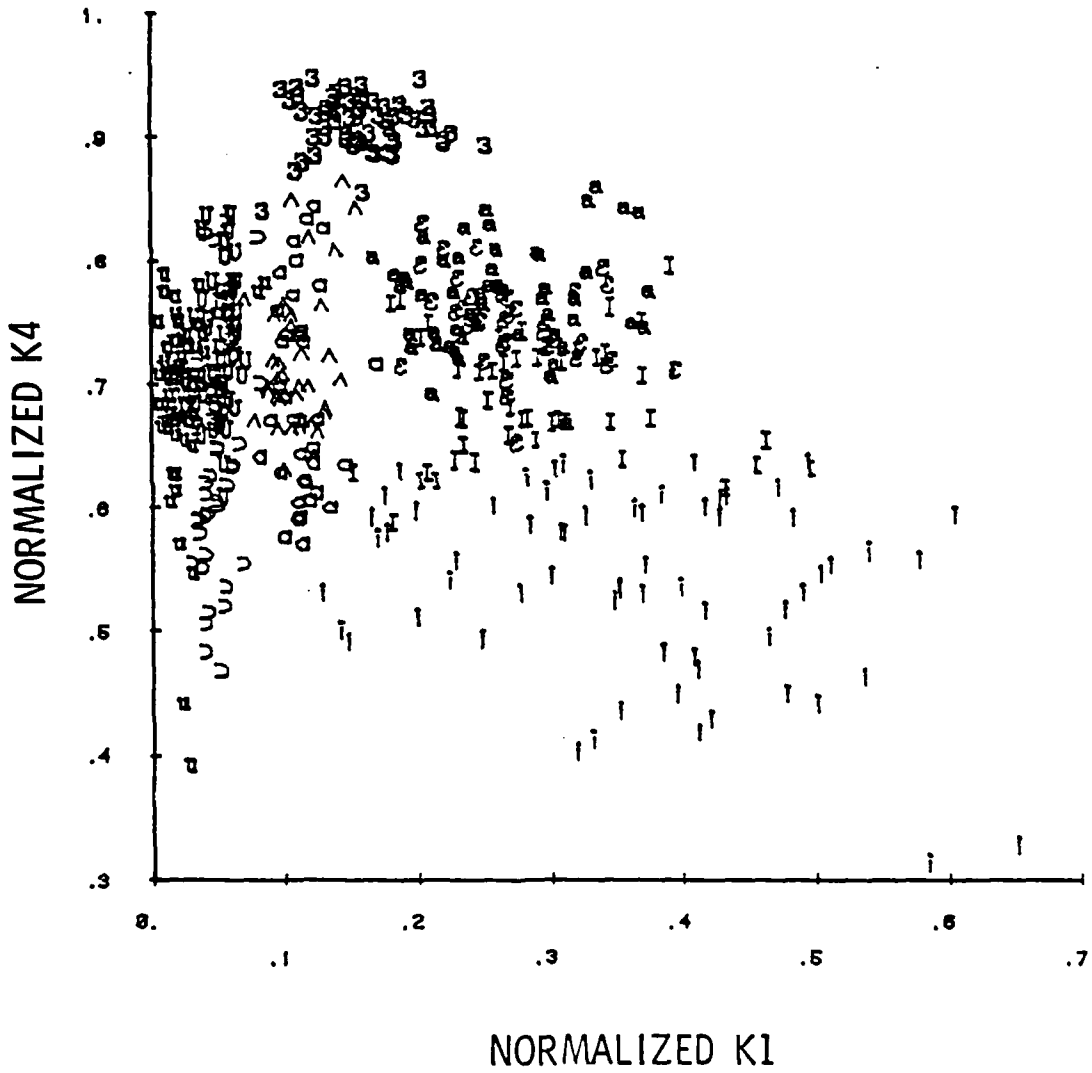


Figure 26. Clustering of the ten English vowels in the K1-K4 plane.

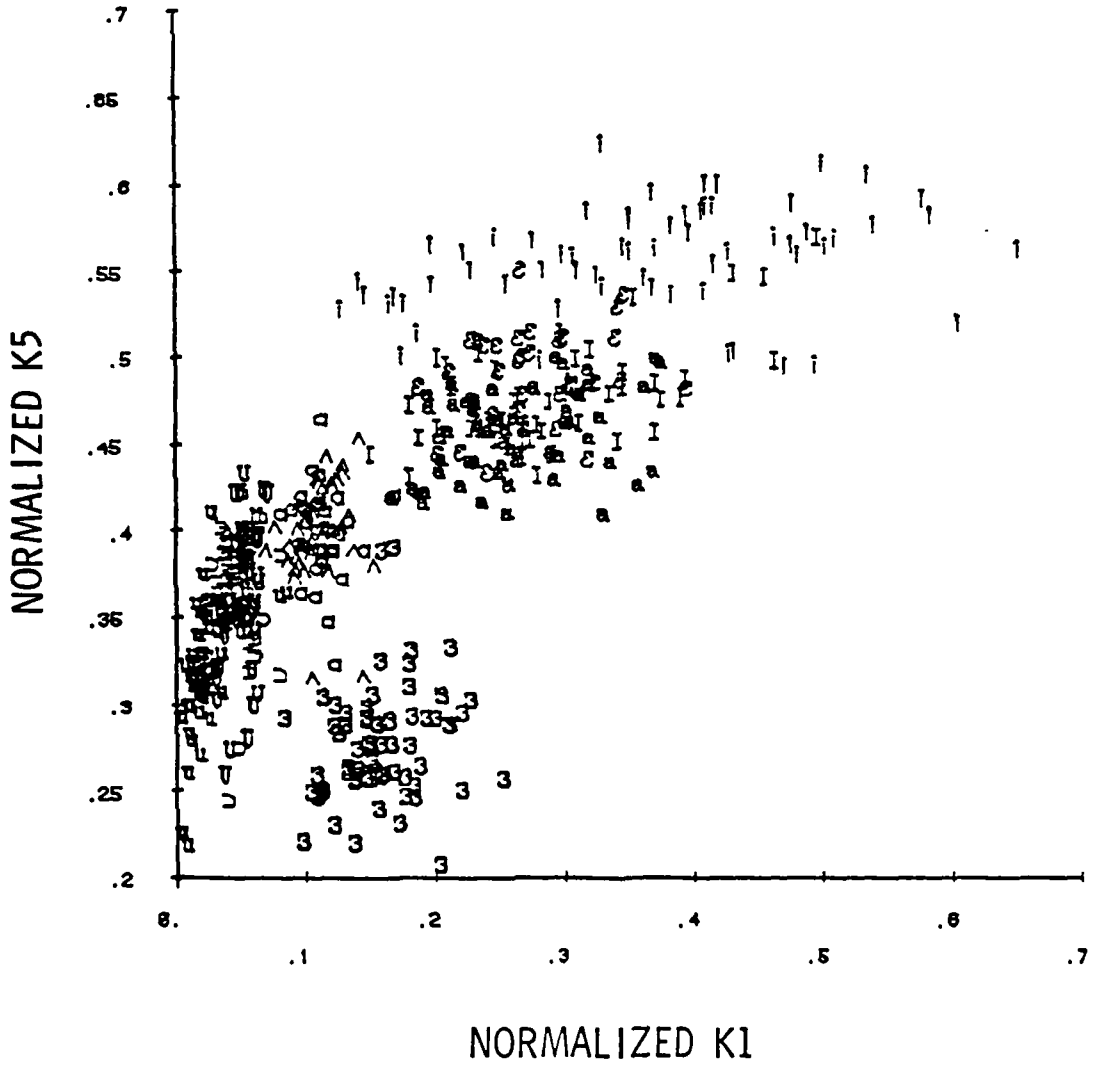


Figure 27. Clustering of the ten English vowels in the K1-K5 plane.

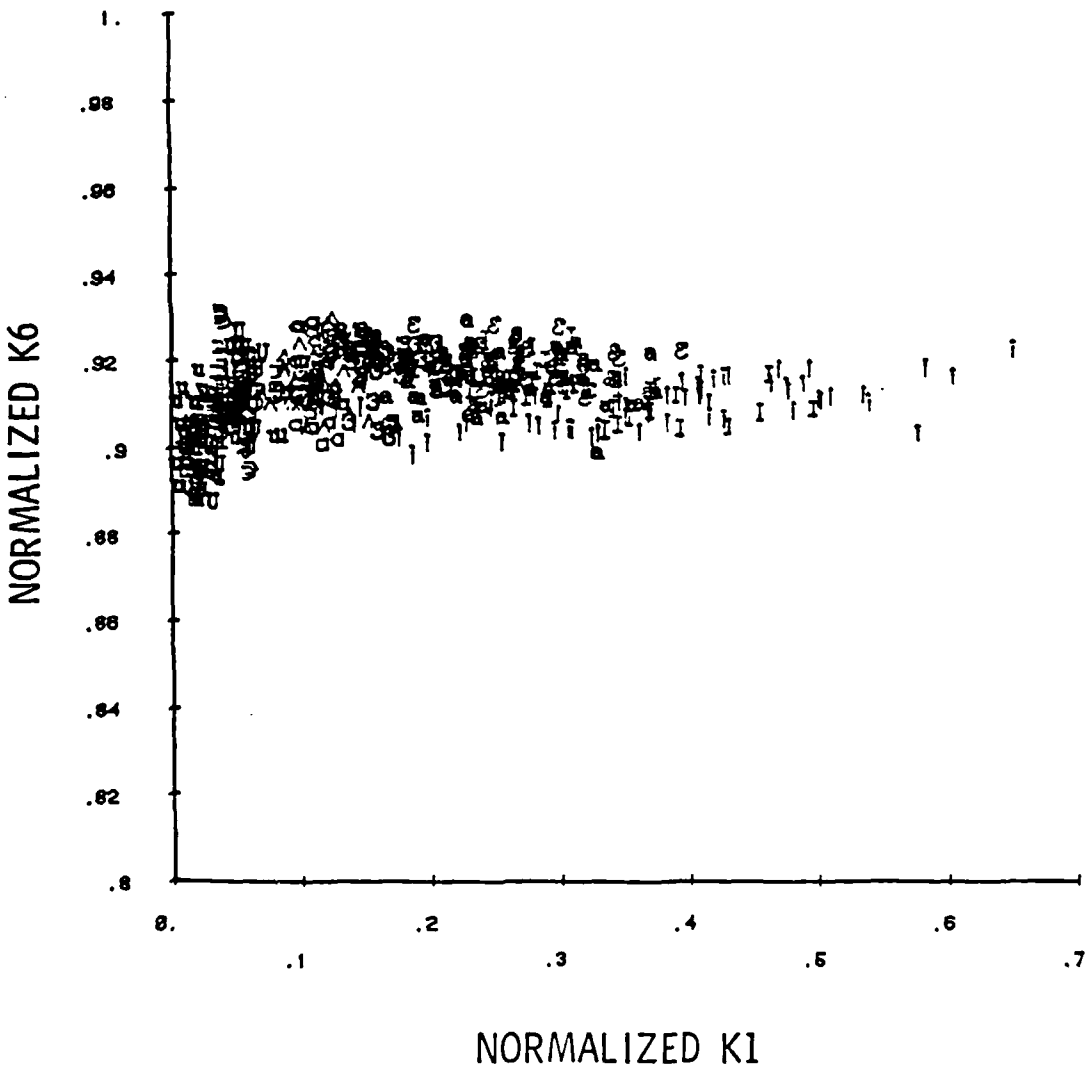


Figure 28. Clustering of the ten English vowels in the K1-K6 plane.

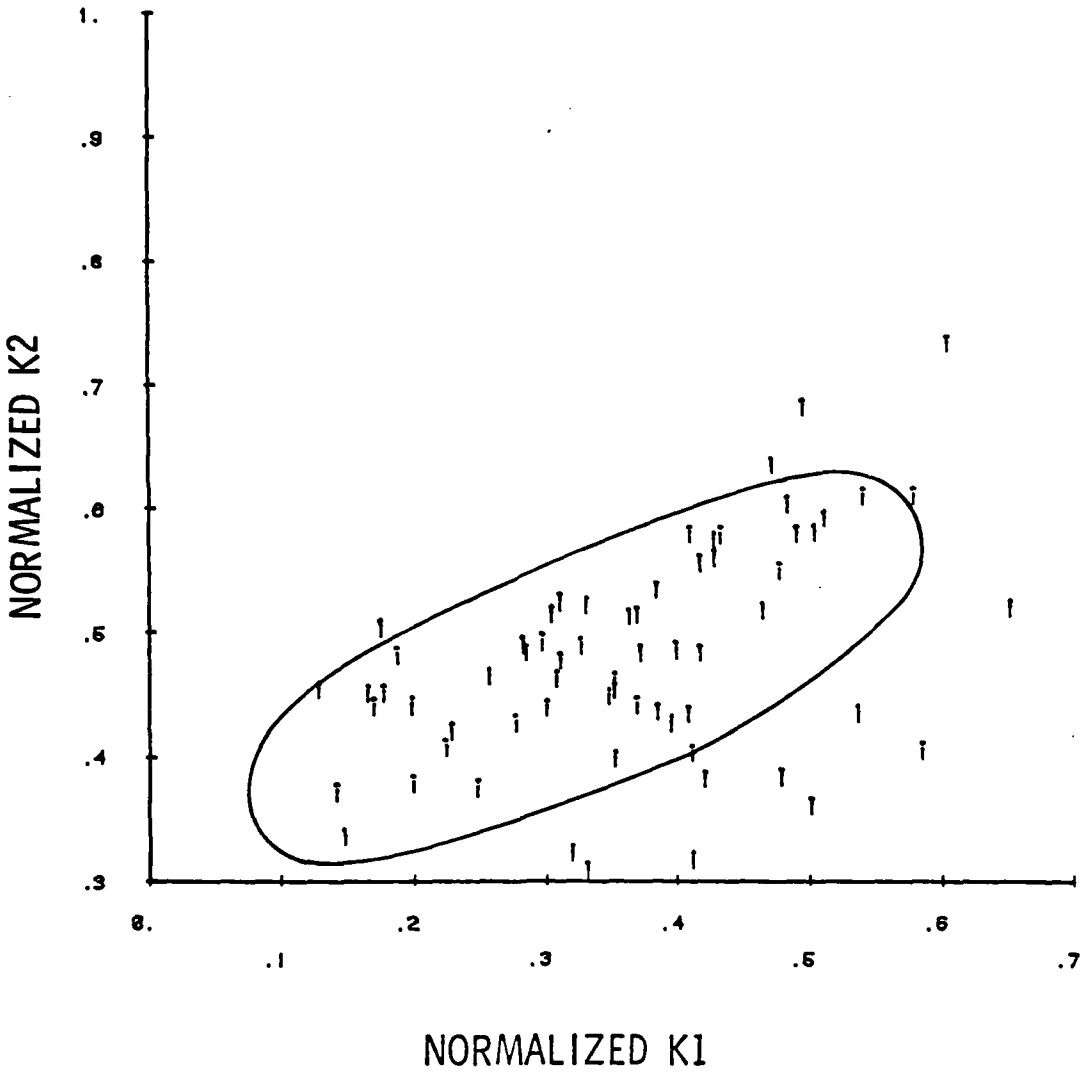


Figure 29. Clustering of the vowel /i/ in the K1-K2 plane.

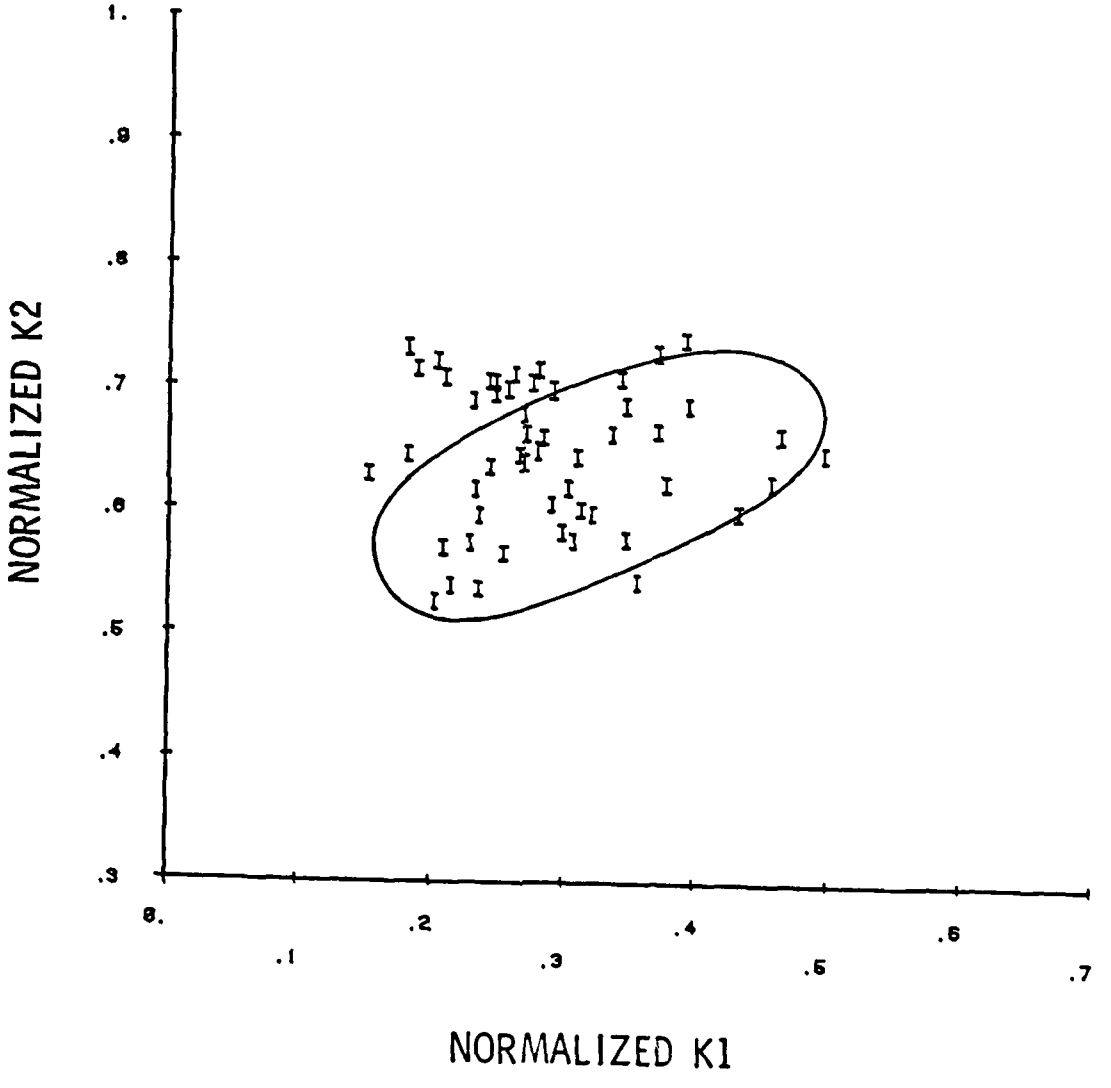


Figure 30. Clustering of the vowel /I/ in the K1-K2 plane.

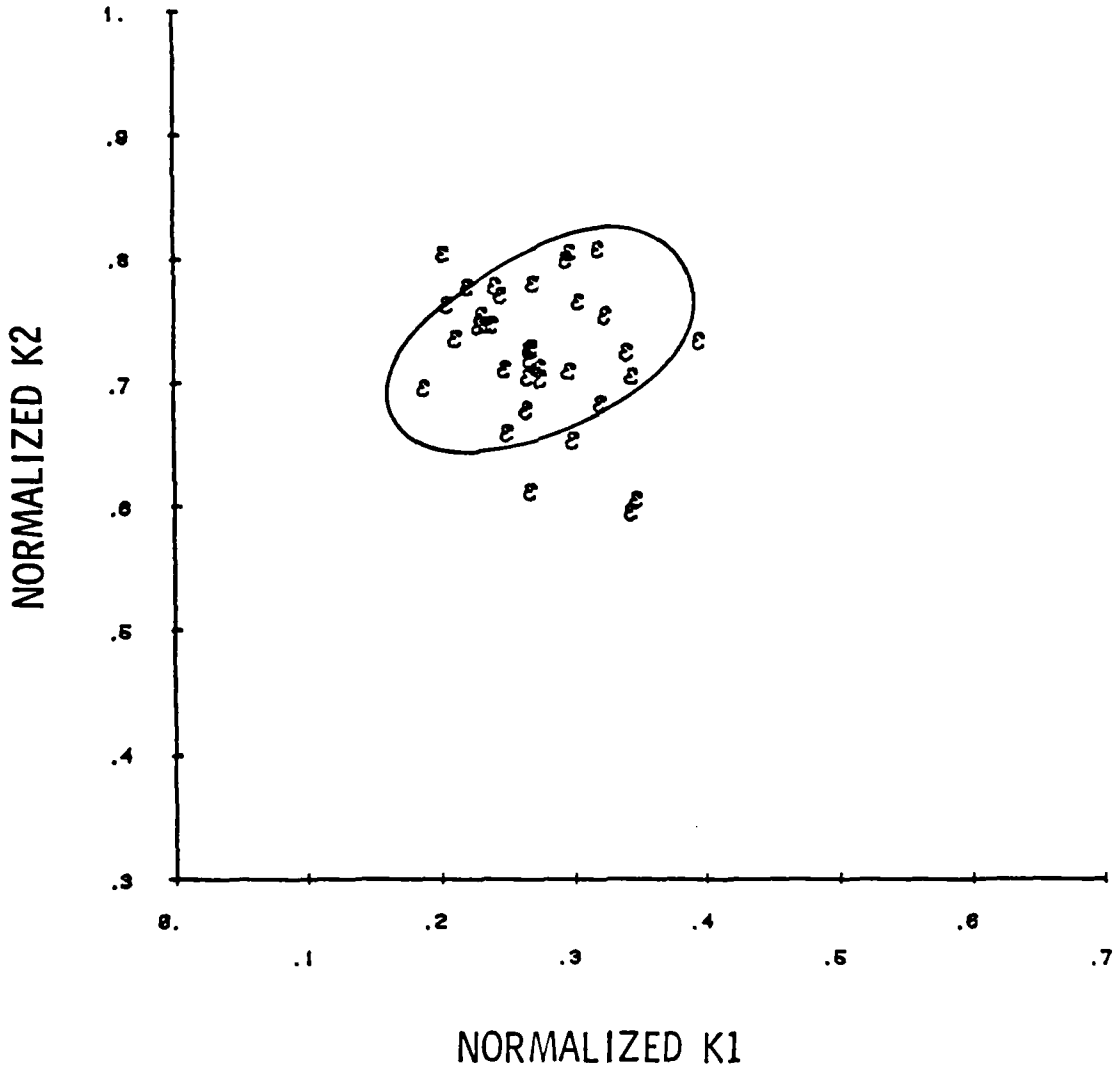


Figure 31. Clustering of the vowel /ɛ/ in the K1-K2 plane.

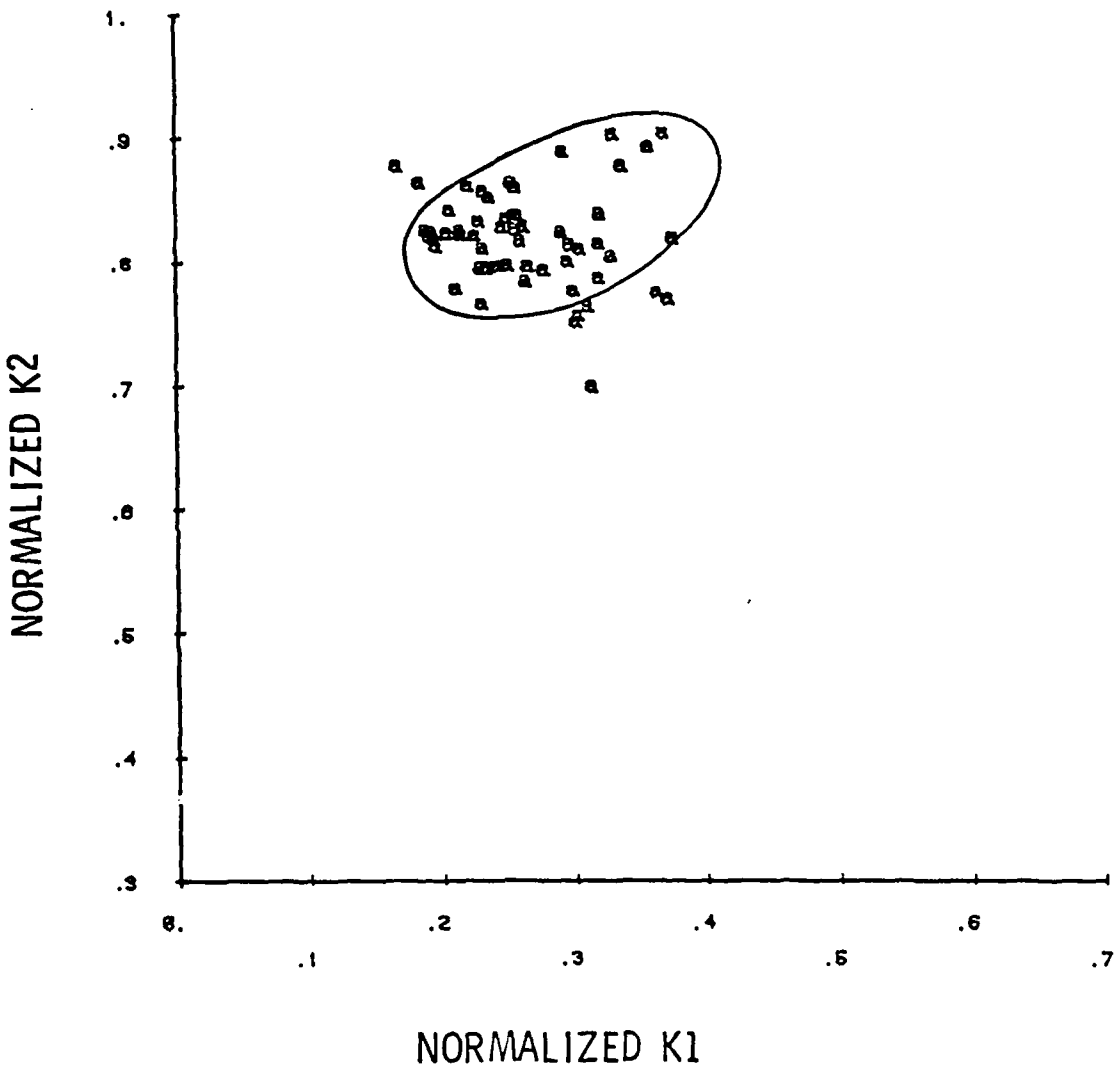


Figure 32. Clustering of the vowel /æ/ in the K1-K2 plane.

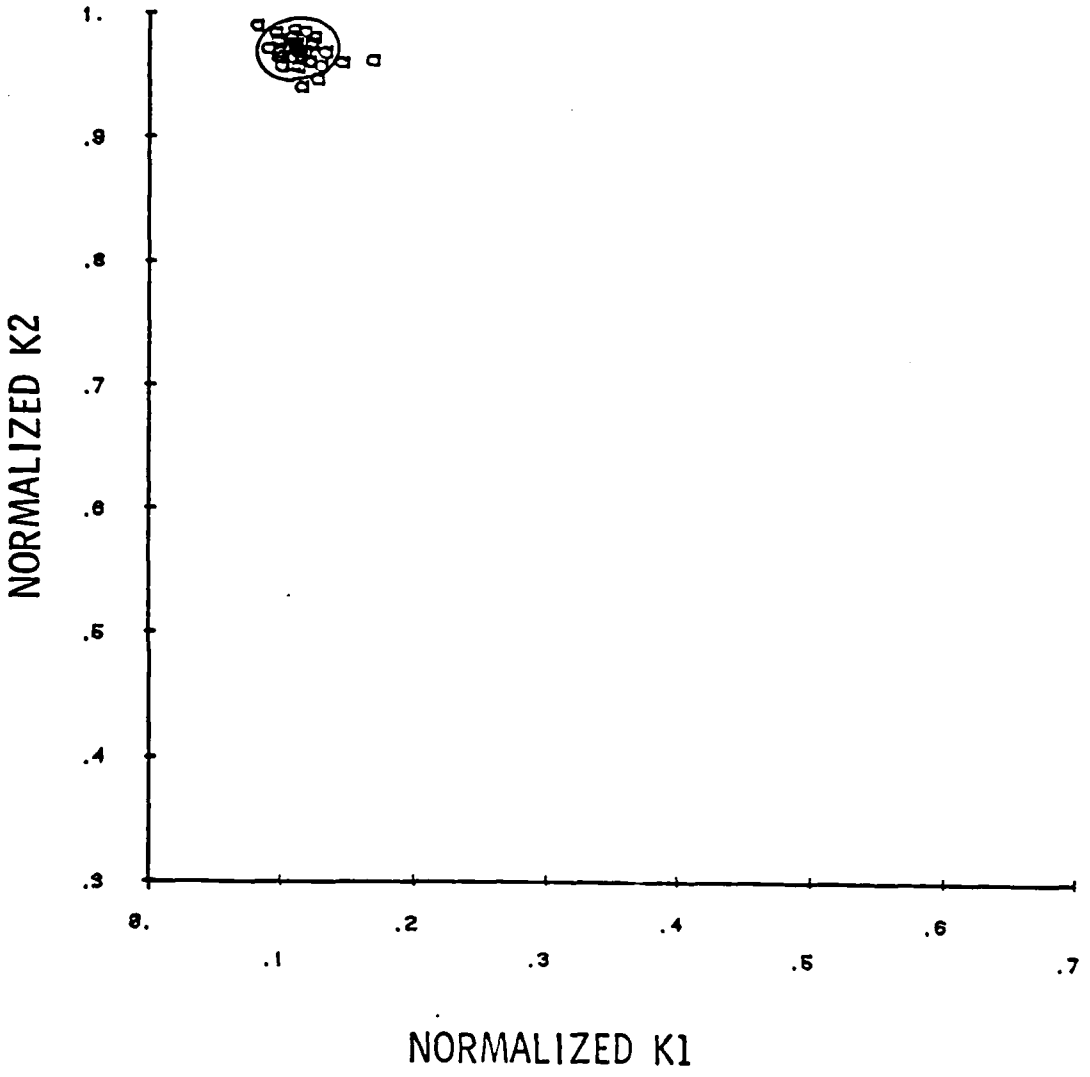


Figure 33. Clustering of the vowel /ɑ/ in the K1-K2 plane.

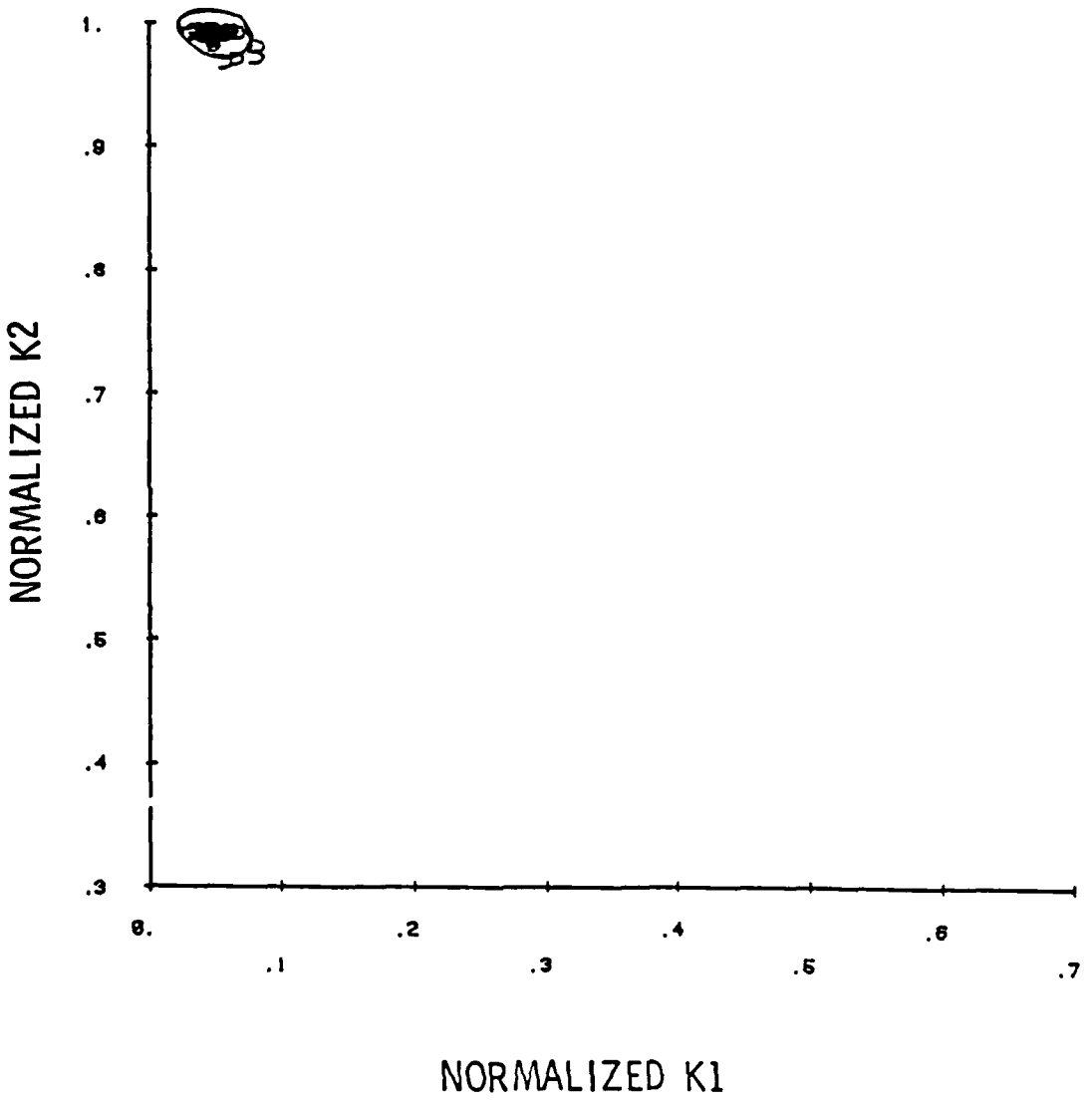


Figure 34. Clustering of the vowel /ɔ/ in the K1-K2 plane.

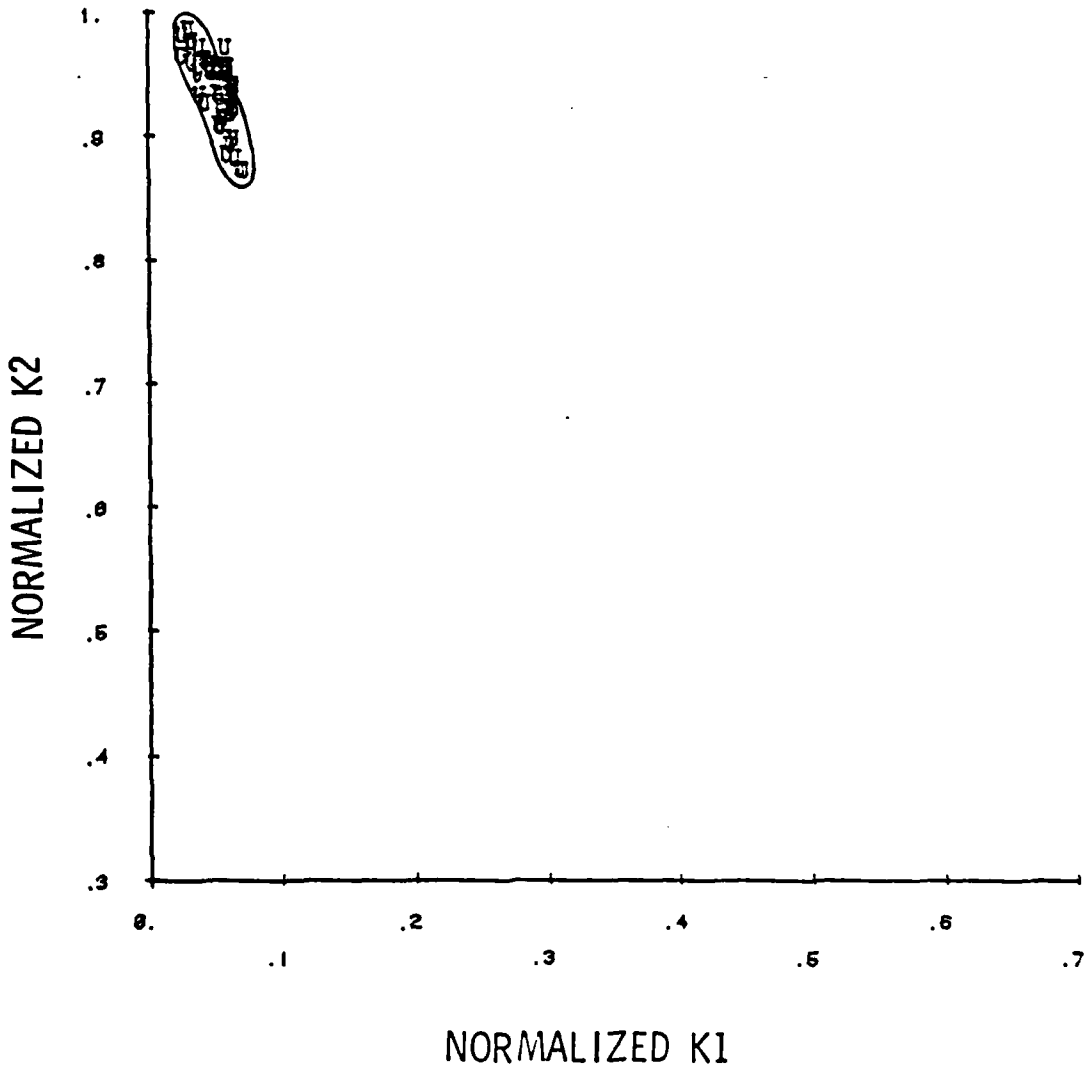


Figure 35. Clustering of the vowel /U/ in the K1-K2 plane.

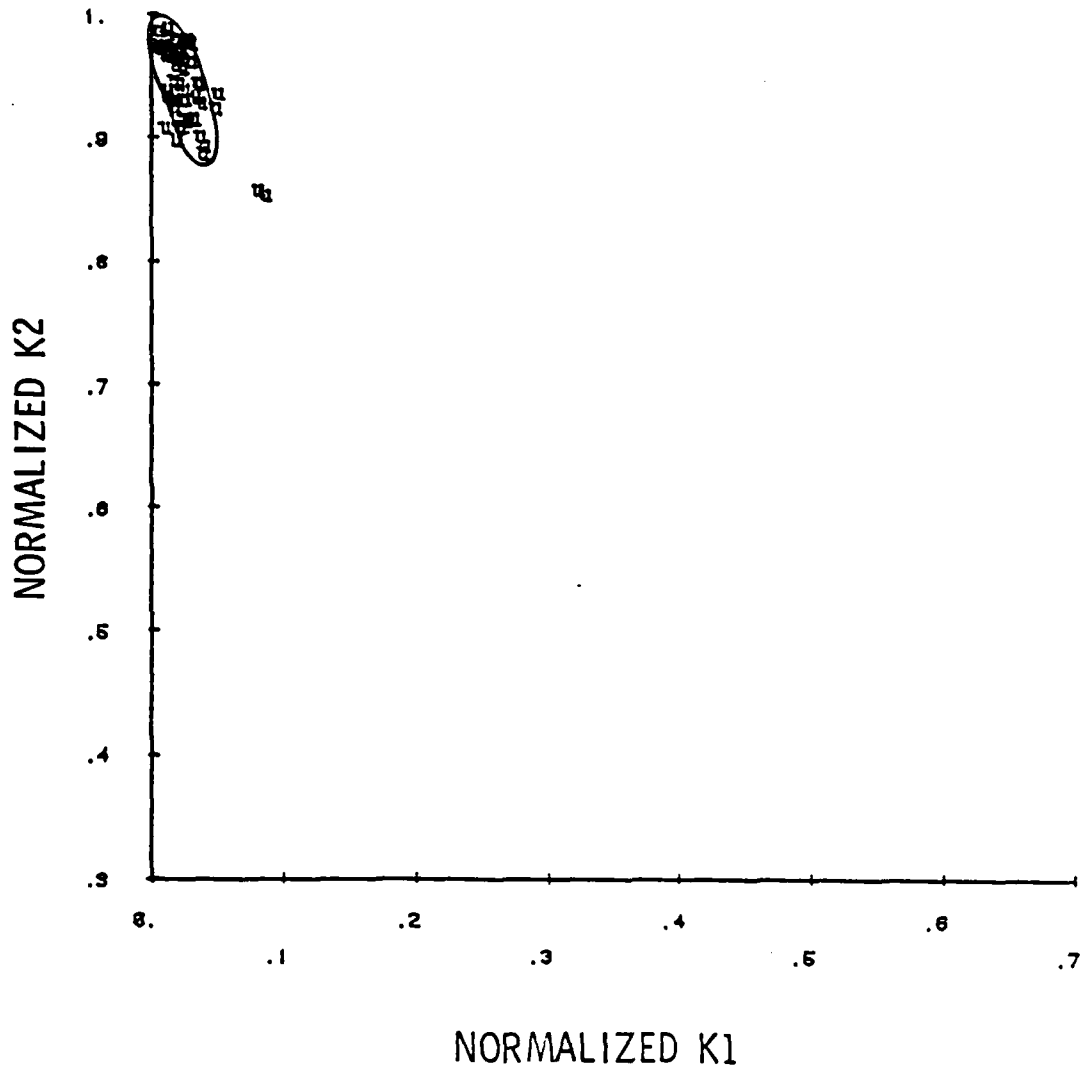


Figure 36. Clustering of the vowel /u/ in the K1-K2 plane.

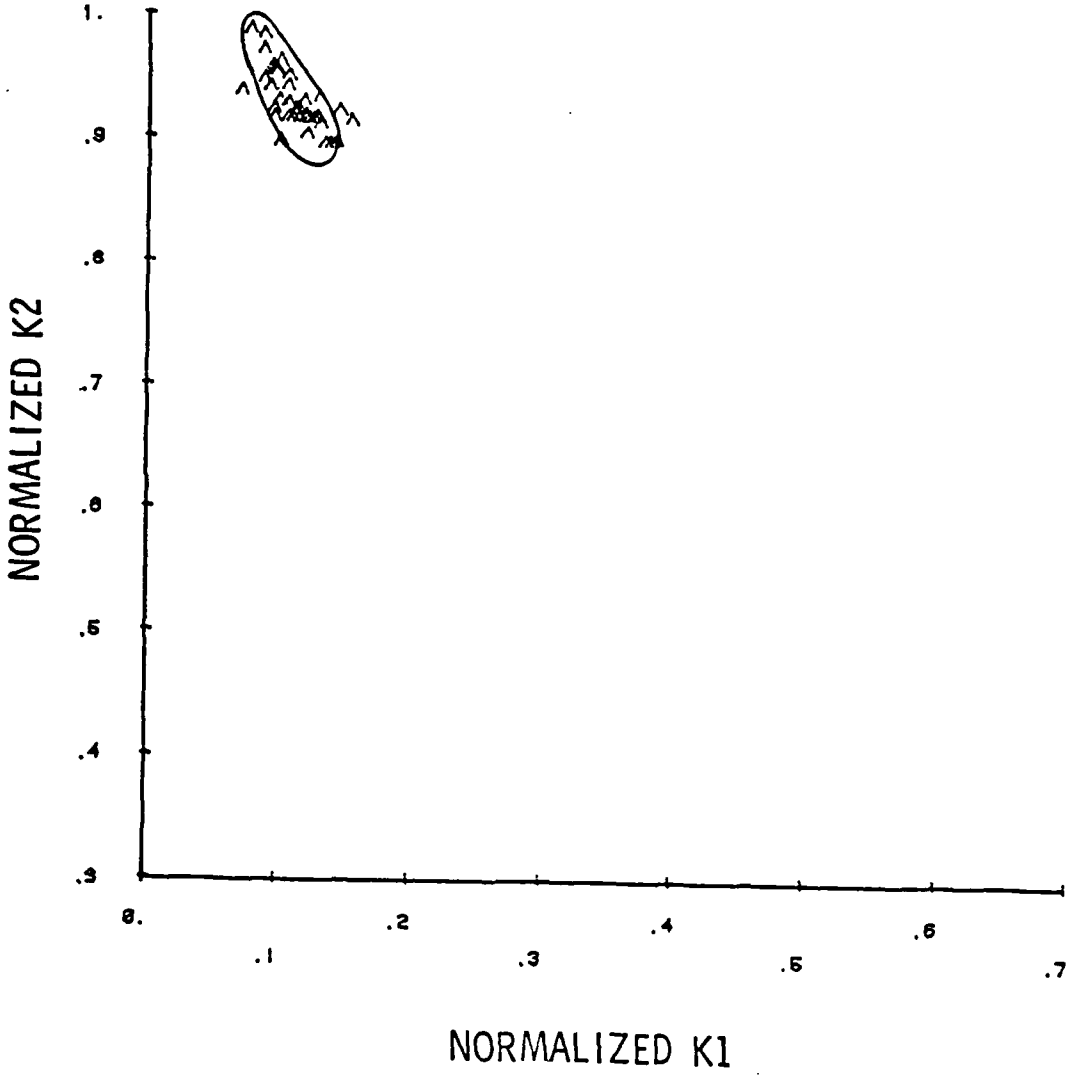


Figure 37. Clustering of the vowel /Λ/ in the K1-K2 plane.

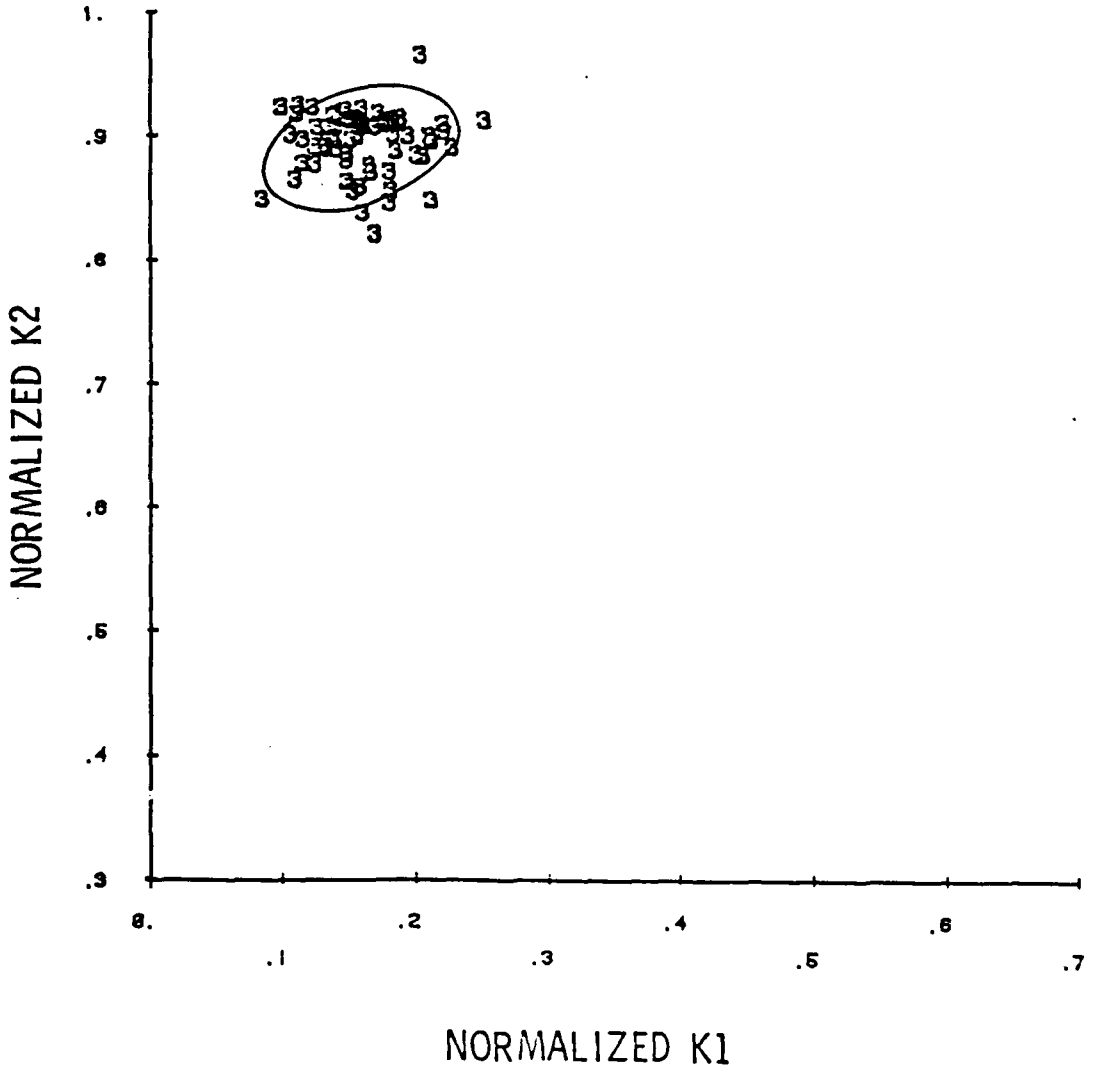


Figure 38. Clustering of the vowel /ɜ/ in the K1-K2 plane.

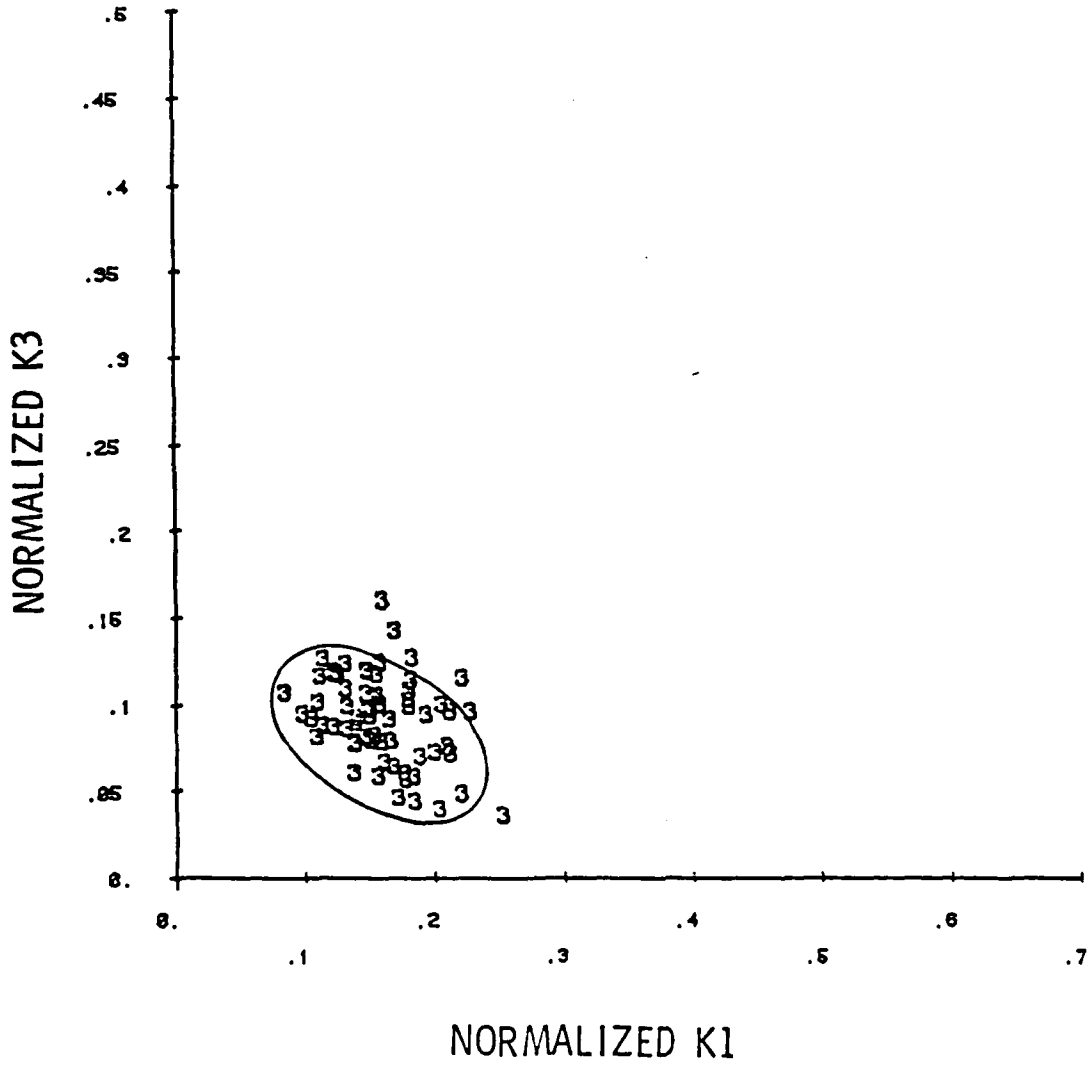


Figure 39. Clustering of the vowel /3/ in the K1-K3 plane.

Distance Measures

The same selected quantitative measures of the cluster sizes and relationships which are computed for formant frequency clusters are also computed in two, three, and six dimensions for the PARCOR coefficient clusters. Table 3 (p. 35) presents the average intracluster distance for each vowel, for two, three, and six-dimensional cases. As for the formant frequencies, intracluster distances are minimum (indicating cluster compactness) for every vowel in two dimensions. Table 4 (p. 36) presents the intercluster distances for selected adjacent vowel pairs for the two, three, and six-dimensional cases. As for the formant frequencies, intercluster distances are maximum (indicating separability) in the largest number of dimensions, which, for the PARCOR coefficients is the six-dimensional case.

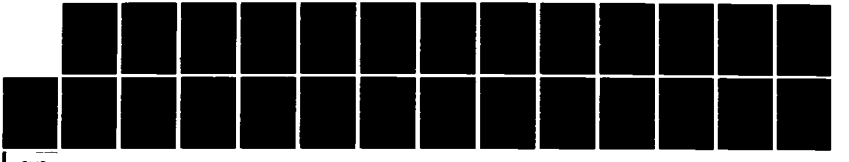
The efficiency with which the PARCOR coefficients represent the vowel clusters may be compared to that exhibited by the formant frequencies by studying Tables 3 and 4 (pp. 35-36). The average intracluster distance (Table 3) should be minimized and the intercluster distances (Table 4) should be maximized.

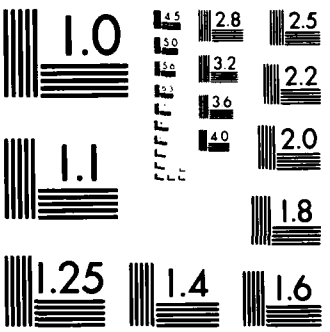
As a combined measure of compactness and separability, The ratio of the sum of average intracluster distances to intercluster distance for adjacent-vowel pairs is computed

for formants and PARCOR coeffients in two, three, and six dimensions. This parameter should be minimized for clusters which are both separable and compact; it is greater than one for clusters which are less so. Quantitatively, analysis of this parameter (Table 5, p. 37) indicates a smaller ratio for PARCOR coefficients than for formants for five of the twelve adjacent-vowel pairs. In other words, the PARCOR coefficient clusters are roughly equivalent the the formant frequency clusters in terms of their compactness and separability. The actual number of dimensions in which the smaller ratios are obtained varies over the vowel pairs. The use of this ratio must be coupled with a qualitative assessment of the vowel clusters. For instance, although the vowels /i/ and /I/ in the K1-K2 plane in Figures 29 and 30 are both widely dispersed, inspection of Figure 24 reveals that the vowel-like sounds may be separated in the K1-K2 plane. The combined ratio for the pair i-I, however, suffers because the coefficients are so widely dispersed. Inspection of the analogous figures (2,4, and 5) for the formant frequencies suggests that the formant space representation is about equivalent to the PARCOR coefficient representation, yet the combined ratio in the formant space is smaller due to the more compact nature of the clusters. The elliptical shapes of the clusters also contribute to the inaccuracy of this type of measurement, since it is more suited to clusters which are symmetric.

AD-A131 503 CLUSTERING OF THE LEAST SQUARES LATTICE PARCOR (PARTIAL CORRELATION) COEF..(U) PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESEARCH LAB.. B A COOPER 2/2

UNCLASSIFIED AUG 83 ARL/PSU/TM-83-90 N00024-79-C-6043 F/G 12/1 NL





An item by item comparison between the columns in each Table is meaningful, specifically the two-dimensional columns in Table 3 and the F(1,2,3) and K(1-6) columns in Table 4. This method of evaluating the two systems of classifying the vowels is more meaningful and informative in conjunction with a graphical assessment since the results are not consistently in favor of one system or the other. The vowel clusters are not consistently more compact or well separated in one domain than in the other. In other words, the PARCOR coefficient representation of the vowels is about equivalent to the formant representation.

It is possible that taking the logarithm of the PARCOR coefficients would cause them to cluster more compactly, since the clusters have an elliptical shape. This is not done, however, because there is no physical justification for the transformation (such as the nonlinearity of the ear in the case of the formant frequencies).

Various distance measures have been employed by researchers in the speech field to assess the similarity between two utterances. These distance measures are commonly computed from the linear predictor coefficients, $\hat{a}_j$, (Atal & Rabiner, 1976; Gray & Markel, 1976; Levinson, Rabiner, Rosenberg & Wilpon, 1979; Tribolet,Rabiner, & Sondhi, 1979) The physical significance of the PARCOR coefficients lends credibility to their use as an

alternative vehicle for assessing similarity between utterances. A quantitative comparison between the various measurements is warranted, based on the results presented here which indicate that the PARCOR coefficients may be equivalent to or better than one or more of the more widely used feature parameters.

CHAPTER VI

SUMMARY AND CONCLUSIONS

It was shown by Potter and Steinberg (1950) and by Peterson and Barney (1952) that the vowel formant frequencies F1 and F2 tend to cluster by vowel when plotted for different speakers. Formant frequency data measured by these researchers for utterances by male speakers were obtained and analyzed by this author quantitatively as well as graphically. The results obtained by the Bell Laboratories researchers (Peterson & Barney, 1952; Potter & Steinberg, 1950) are verified; the first nine vowel clusters are defined in the space defined by F1 and F2, whereas the third formant is necessary for identification of the tenth vowel, /ɜ/. The average intracluster distance for each vowel cluster yields smaller values for each vowel when computed in two dimensions rather than three, indicating that the clusters are most compact in two dimensions. Intercluster distances between adjacent vowel pairs were computed as a measure of vowel separability and found to be maximum in three dimensions for all of the pairs.

Each of the utterances was then reproduced from the formant frequencies as an autoregressive time series by a six-pole IIR recursive digital filter. These time series were then inverse filtered with a six-zero complex adaptive

lattice filter adapted from Alexandrou & Hodgkiss (Note 1) which yielded a whitened output signal. The partial correlation (PARCOR) coefficients from this lattice filter were shown to cluster by vowel in the space defined by these coefficients.

Graphically, the first two coefficients, K1 and K2, are sufficient to identify the first nine vowels, whereas the third PARCOR coefficient, K3, is necessary to distinguish the tenth vowel, /ɜ/, from the other nine. The results of a numerical analysis of the PARCOR coefficients were analagous to those found for the formant frequencies. From calculations of average intracluster distance for each vowel, it was determined that the clusters are most compact in two dimensions (K1,K2). Calculations of intercluster distance between adjacent-vowel pairs show maximum cluster separation in the largest number of dimensions (six). The ratio of the sum of intracluster distances to intercluster distance for each of the adjacent-vowel pairs indicates that the PARCOR coefficient representation is as effective or better than the formant frequency representation for five of the twelve adjacent-vowel pairs. It is apparent then, that the use of the PARCOR coefficients for identification of synthesized steady state vowel-like sounds is as effective as identification via formant frequencies. The PARCOR coefficient technique for the identification of steady state

synthesized vowel-like sounds is a much quicker and more computationally efficient method than that involving computation of poles and zeros and back calculation of formant frequencies and bandwidths. This is very important in real-time identification of non-stationary signals.

Limitations of the Study

The least squares lattice is an optimal whitening filter for an AR process when the order of the lattice (number of zeros) is equal to the order of the AR process (number of poles). For this study this is the case, as it is desired to obtain the PARCOR coefficients for inputs of known order. However, for an input signal whose origin is not known, the performance of the filter will depend highly on the order which is selected.

Another simplification made in this study for the purpose of exactly matching the input and output transfer functions is the modeling of the speech signal as an AR process. This is a commonly used representation in the literature, although it is extremely simplified. The assumption of an AR process may only be made for non-nasal sounds because the coupling of nasal cavities during production of nasalized sounds adds an antiresonance or zero to the speeech spectrum (Denes & Pinson, 1963). The process may no longer be accurately modeled as all-pole. However,

researchers have commonly used an all-pole representation of higher order (Atal & Hanauer, 1971; Friedlander, 1982; Kay & Marple, 1981; Rabiner & Schafer, 1978) for this purpose as an approximation to a more desirable pole-zero (ARMA) model because AR models are much easier to use.

The most severe limitation of the study is the fact that synthesized speech-like sounds (rather than actual speech sounds) are used. Although synthesized sounds are used intentionally for the specific purpose of establishing the PARCOR coefficients as equivalent to formant frequencies as pattern recognition features, further studies need to concentrate efforts on identifying actual spoken speech. Actual speech cannot be accurately modeled as AR (even steady-state vowel sounds) because the spectrum will contain extra poles and zeros which are contributed by the following factors: higher formant frequencies, lip radiation, actual (not flat) excitation spectrum, aperiodicity of the excitation function, damping of the vocal tract, any laryngeal pathologies, measurement difficulty and error, transmission loss between lips and microphone, and inaccuracies in the mathematical speech production model (Dunn, 1961; Fant, 1956,1959, 1963; Peterson, 1959; Rabiner & Schafer, 1978, chap.3). An ARMA lattice is appropriate in the case of actual speech in order to more accurately estimate the spectrum. Friedlander and Mitra

(1981) have used an ARMA lattice for the identification of actual spoken nasal sounds; the results compared favorably with those obtained by using a high order AR lattice. This is also discussed by Fallside and Brooks (1976), Green (1976), and Markel and Gray (1976). ARMA lattice algorithms are presented by Lee, Friedlander and Morf, (1980), Morf, Lee, Nickolls, and Vieira (1977), and Morf, Vieira and Lee (1977).

Suggestions for Future Research

Very few speech sounds are steady state, and only very briefly if at all. Use of the adaptive capability of the lattice filter with appropriate selection of the fade factor, $(1-\alpha_{CLSL})$, will enable the results of vowel identification studies to be extended to simplify the identification of more complex time-varying sounds. Dipthongs are commonly identified by researchers in the speech field (Potter, Kopp & Kopp, 1961) by the time-varying paths of their second formant frequencies between formant locations for the two composite vowel sounds. It seems reasonable that this could be transformed into an identification via time-varying PARCOR coefficient(s). Turner (1982) has used the time-varying behavior of the PARCOR coefficients to identify stop consonants, which are characterized by an even more complicated time-varying frequency spectrum. It is likely that with extensions to a

higher order AR or ARMA lattice the technique of using the PARCOR coefficients as pattern recognition features would be well suited to a multitude of applications in the field of signal processing. Whether used to identify stationary signals or to adaptively identify and track any type of acoustic signal, the PARCOR coefficients of the complex adaptive least squares lattice conveniently and efficiently represent time domain signals. In a purely pattern recognition context, the PARCOR coefficients are valuable pattern recognition features in situations where the frequency spectrum or pole locations are meaningless. Deller and Anderson(1980) identified types of laryngeal pathologies by looking at clusters of z-plane pole locations in and on the unit circle. The actual pole locations have a complicated relationship to the actual pathology; in their case, all that was needed was a clustering parameter to identify outlying points and types of clusters.

The identification of synthesized steady state vowel-like sounds is a first step in the process of speech identification. The clustering properties of the PARCOR coefficients which are demonstrated in this research for the purpose of vowel identification show the PARCOR coefficients to be an effective and efficient vehicle for the representation and transmission of frequency spectra information. It is hoped that these results will inspire

other researchers to extend the study to enable the simplification of other more complex system identification problems.

REFERENCE NOTES

1. Alexandrou, D., & Hodgkiss, W. S. Personal communication, 23 June 1982.

2. Barney, H. L. Fundamental and formant frequencies of a group of vowel sounds classified by a sub-jury of 26 observers--case 38138-2. Short Hills, N.J.: Bell Laboratories, Inc., 2 February 1951.

3. Hodgkiss, W. S. Personal communication, 8 July 1982.

4. Pack, J. D., & Satorius, E. H. Least squares, adaptive lattice algorithms (NOSC TR 423). San Diego: Naval Ocean Systems Center, April 1979.

REFERENCES

Atal, B. S., & Hanauer, S. L. Speech analysis by linear prediction of the speech wave. Journal of the Acoustical Society of America, 1971, 50(2), 637-655.

Atal, B. S., & Rabiner, L. R. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, ASSP-24(3), 201-212.

Bernstein, J. Formant-based representations of auditory similarity among vowel-like sounds. Journal of the Acoustical Society of America, 1981, 69(4), 1132-1144.

Bogert, B. P. On the band width of vowel formants. Journal of the Acoustical Society of America, 1953, 25(4), 791-792.

Burg, J. P. Maximum entropy spectral analysis. Proceedings of the 37th Meeting of the Society of Exploration Geophysicists, 1967, 34-40.

Deller Jr., J. R., & Anderson, D. J. Automatic classification of laryngeal dysfunction using the roots of the digital inverse filter. IEEE Transactions on Biomedical Engineering, 1980, BME-27(12), 714-721.

Denes, P. B., & Pinson, E. N. The Speech Chain. Bell Telephone Laboratories, 1963.

Dunn, H. K. Methods of measuring vowel formant bandwidths. Journal of the Acoustical Society of America, 1961, 33(12), 1737-1746.

Fallside, F., & Brooks, S. Analysis and areas modelling of nasalised speech by a multivariable identification technique. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1976, 80. (Abstract)

Fant, G., Fintoft, K., Liljencrants, J., Lindblom, B., & Martony, J. Formant amplitude measurements. Proceedings of the Speech Communication Seminar, 1963, 1-13.

Fant, G. The acoustics of speech. Proceedings of the Third International Congress on Acoustics, 1959, 188-201.

Fant, C. G. M. On the predictability of formant levels and spectrum envelopes from formant frequencies. In Halle, M., Lunt, H. G., McLean, H., & Van Schooneveld, C. H. (Eds.), For Roman Jakobson. The Hague: Mouton ('s-Gravenhage), 1956.

Flanagan, J. A. Speech Analysis, Synthesis, and Perception (2nd ed.). New York: Springer, 1972.

Foulkes, J. D. Computer identification of vowel types. Journal of the Acoustical Society of America, 1961, 33(1), 7-11.

Friedlander, B. Recursive lattice forms for spectral estimation and adaptive control. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1980, 466-471.

Friedlander, B. Lattice filters for adaptive processing. Proceedings of the IEEE, 1982, 70(8), 829-867. (a)

Friedlander, B. System identification techniques for adaptive signal processing. Circuits, Systems, and Signal Processing, 1982, 1(1), 3-41. (b)

Friedlander, B., & Mitra, S. Speech deconvolution by recursive ARMA lattice filters. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1981, 343-346.

Gray, A. H., & Markel, J. D. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, ASSP-24(5), 380-391.

Green, N. Analysis-synthesis using pole-zero approximations to speech spectra. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1976, 306-309.

Griffiths, L. J. Rapid measurement of digital instantaneous frequency. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1975, ASSP-23(2), 207-222.

Griffiths, L. J. A continuously adaptive filter implemented as a lattice structure. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1977, 683-686.

Griffiths, L. J. An adaptive lattice structure for noise-cancelling applications. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1978, 87-90.

Hodgkiss,W. S., & Alexandrou, D. Application of adaptive linear predictor structures to the prewhitening of acoustic reverberation data. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1983, 599-602.

Hodgkiss, W. S., & Presley Jr., J. A. Adaptive tracking of multiple sinusoids whose power levels are widely separated. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, ASSP-29(3), 710-721.

Hodgkiss, W. S., & Presley Jr., J. A. The complex adaptive least-squares lattice. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1982, ASSP-30(2), 330-333.

Itakura, F., & Saito, S. Digital filtering techniques for speech analysis and synthesis. Proceedings of the 7th International Congress on Acoustics, 1971, 261-264.

Kay, S. M., & Marple, Jr., S. L. Spectrum analysis--a modern perspective. Proceedings of the IEEE, 1981, 69(11), 1380-1419.

Landers, T. E., & Lacoss, R. T. Some geophysical applications of autoregressive spectral estimates. IEEE Transactions on Geoscience Electronics, 1977, GE-15(1), 26-32.

Lee, D. T. L. Canonical ladder form realizations and fast estimation algorithms (Doctoral dissertation, Stanford University, 1980). Dissertation Abstracts International, 1980, 41, 3126B. (University Microfilms No. 81-03530)

Lee, D. T. L., Friedlander, B., & Morf, M. Recursive ladder algorithms for ARMA modeling. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1980, 1225-1231.

Lee, D. T. L., Morf, M., & Friedlander, B. Recursive least squares ladder estimation algorithms. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, ASSP-29(3), 627-641.

Levinson, N. The Wiener RMS (Root-mean square) error criterion in filter design and prediction. Journal of Mathematics and Physics, 1947, 25(4), 261-278.

Levinson, S. E., Rabiner, L. R., Rosenberg, A. E., & Wilpon, J. G. Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, ASSP-27(2), 134-140.

Macina, N. A. Maximum entropy modeling for identification and detection of certain classes of dynamic events. RCA Review, 1981, 42, 85-110.

Makhoul, J. Linear prediction: A tutorial review. Proceedings of the IEEE, 1975, 63(4), 561-580.

Makhoul, J. A class of all-zero lattice digital filters: properties and applications. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, ASSP-26(4), 304-314.

Markel, J. D. Digital inverse filtering--a new tool for formant trajectory estimation. IEEE Transactions on Audio and Electroacoustics, 1972, AU-20(2), 129-137.

Markel, J. D. Application of a digital inverse filter for automatic formant and Fo analysis. IEEE Transactions on Audio and Electroacoustics, 1973, AU-21(3), 154-160.

Markel, J. D., & Gray, A. H. Linear Prediction of Speech. New York: Springer, 1976.

McCandless, S. S. An algorithm for automatic formant extraction using linear prediction spectra. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1974, ASSP-22(2), 135-141.

Morf, M., Dickenson, B., Kailath, T., & Vieira, A. Efficient solutions of covariance equations for linear prediction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1977, ASSP-25(5), 429-435.

Morf, M., & Lee, D. T. L. Ladder forms for identification and speech processing. Proceedings of the IEEE Conference on Decision and Control, 1978, 1362-1367.

Morf, M., Lee, D. T. L., Nickolls, J. R., & Vieira, A. A classification of algorithms for ARMA models and ladder realizations. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1977, 13-19.

Morf, M., Vieira, A., & Lee, D. T. L. Ladder forms for identification and speech processing. Proceedings of the IEEE Conference on Decision and Control, 1977, 1074-1078.

Papoulis, A. Maximum entropy and spectral estimation: A review. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, ASSP-29(6), 1176-1186.

Parzen, E. Some recent advances in time series modeling. IEEE Transactions on Automatic Control, 1974, AC-19(6), 723-730.

Peterson, G.E. The information-bearing elements of speech. Journal of the Acoustical Society of America, 1952, 24(6), 629-637.

Peterson, G. E. Vowel formant measurements. Journal of Speech and Hearing Research, 1959, 2(2), 173-183.

Peterson, G. E. Parameters of vowel quality. Journal of Speech and Hearing Research, 1961, 4(1), 10-29.

Peterson, G. E. & Barney, H. L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 1952, 24(2), 175-184.

Potter, R. K., Kopp, G. A., & Kopp, H. G. Visible Speech. New York: Dover, 1966.

Potter, R. K., & Steinberg, J. C. Toward the specification of speech. Journal of the Acoustical Society of America, 1950, 22(6), 807-820.

Rabiner, L. R., & Schafer, R. W. Digital Processing of Speech Signals. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Robinson, E.A., & Treitel, S. Maximum entropy and the relationship of the partial autocorrelation to the reflection coefficients of a layered system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, ASSP-28(2), 224-235.

Satorius, E. H., & Alexander, S. T. Channel equalization using adaptive lattice algorithms. IEEE Transactions on Communications, 1979, COM-27(6), 899-905.

Satorius, E. H., & Pack, J. D. Application of least squares lattice algorithms to adaptive equalization. IEEE Transactions on Communication, 1981, COM-29(2), 136-142.

Satorius, E. H., & Shensa, M. J. On the application of recursive least squares methods to adaptive processing. In Narenda & Monopoli (Eds.), Applications of Adaptive Control. New york: Academic Press, 1980. (a)

Satorius, E. H., & Shensa, M. J. Recursive lattice filters--a brief overview. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1980, 955-959. (b)

Tohkura, Y., & Itakura, F. Spectral sensitivity analysis of PARCOR parameters for speech data compression. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, ASSP-27(3), 273-280.

Tou, J. T., & Gonzales, R. C. Pattern Recognition Principles. Reading, Ma.: Addison-Wesley, 1974.

Tribolet, J. M., Rabiner, L. R., & Sondhi, M. M. Statistical properties of an LPC distance measure. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, ASSP-27(5), 550-558.

Turner, J. M. Application of recursive exact least square ladder estimation algorithm for speech recognition. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1982, 543-545.

Ulrych, T. J., & Bishop, T. N. Maximum entropy spectral analysis and autoregressive decomposition. Review of Geophysics and Space Physics, 1975, 13(2), 183-200.

Wakita, H. Direct determination of input impedance singularities from speech for obtaining the vocal tract area. Journal of the Acoustical Society of America, 1973, 53(1), 293-294. (Abstract) (a)

Wakita, H. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Transactions on Audio and Electroacoustics, 1973, AU-21(5), 417-427. (b)

Wakita, H. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, ASSP-27(3), 281-285.

Wakita, H., & Gray, Jr., A. Numerical determination of the lip impedance and vocal tract area functions. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1975, ASSP-23(6), 574-580.

Wakita, H., & Kasuya, H. A study of vowel normalization and identification on connected speech. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1977, 648-651.

Wiggins, R. A., & Robinson, E. A. Recursive solution to the multichannel filtering problem. Journal of Geophysical Research, 1965, 70(8), 1885-1891.

Wood, L., & Treitel, S. Seismic signal processing.

Proceedings of the IEEE, 1975, 63(4), 649-661.

APPENDIX A

Selected Measures of Vowel Cluster Size and Vowel Cluster Separability[a]

Given N_c clusters of p-dimensional pattern vectors $\bar{x}_{kj}$ where x_{kj}^n $n=,2, \ldots p$ is the nth component of that vector and N_j is the number of points in the jth cluster, the centroid vector of the jth cluster C_j is

$$\bar{z}_j = \begin{pmatrix} z_j^1 \\ z_j^2 \\ . \\ . \\ z_j^n \end{pmatrix} = \frac{1}{N_j} \sum_{\bar{x}_{kj} \in C_j} \bar{x}_{kj} \qquad \begin{matrix} k=1,2, \ldots N_j \\ j=1,2, \ldots N_c \end{matrix} .$$

The intercluster distance between the ith and jth cluster is the Euclidean distance between the centroids of the clusters:

$$D_{ij} = ||\bar{z}_i - \bar{z}_j|| = \left(\sum_{n=1}^{p} |z_i^n - z_j^n|^2 \right)^{1/2} = \left[(\bar{z}_i - \bar{z}_j)' (\bar{z}_i - \bar{z}_j) \right]^{1/2} .$$

The intracluster distance is the Euclidean distance from the kth vector in the jth cluster to the mean of that cluster:

$$D_{kj} = ||\bar{x}_{kj} - \bar{z}_j|| \qquad \begin{matrix} k=1,2, \ldots N_j \\ j=1,2, \ldots N_c \end{matrix} .$$

The average intracluster distance for the jth cluster is

$$D_j = \frac{1}{N_j} \sum_{k=1}^{N_j} D_{kj} \ .$$

[a] from Tou and Gonzales (1974, p. 77)

APPENDIX B

Least Squares Lattice Equations[a]

Initialization ($i=0,1,\ldots p$)

$$r_i(-1)=0, \qquad i\neq p \tag{4a}$$

$$E_i^r(-1)=\epsilon_{CLSL}, \qquad \epsilon_{CLSL}=0.001 \text{ and } i\neq p \tag{4b}$$

$$\Delta_i(-1)=0, \qquad i\neq 0 \tag{4c}$$

$$b_k^{(i)}(-1)=0, \qquad 0\leq k\leq i-1, i\neq 0, \text{ and } i\neq p. \tag{4d}$$

Time update ($n\geq 0$)

$$e_0(n)=r_0(n)=x(n) \tag{4e}$$

$$E_0^e(n)=E_0^r(n)=(1-\alpha_{CLSL})E_0^r(n-1)+|x(n)|^2 \tag{4f}$$

$$\gamma_{-1}(n-1)=0. \tag{4g}$$

Order update ($i=1,2,\ldots,p$).

$$\Delta_i(n)=(1-\alpha_{CLSL})\Delta_i(n-1)-\frac{e_{i-1}(n)-r_{i-1}^*(n-1)}{1-\gamma_{i-2}(n-1)} \tag{4h}$$

$$K_i^e(n)=\Delta_i^*(n)/E_{i-1}^e(n) \tag{4i}$$

$$K_i^r(n)=\Delta_i(n)/E_{i-1}^r(n-1) \tag{4j}$$

$$e_i(n)=e_{i-1}(n)+K_i^r(n)r_{i-1}(n-1) \tag{4k}$$

$$r_i(n)=r_{i-1}(n-1)+K_i^e(n)e_{i-1}(n) \tag{4l}$$

$$E_i^e(n)=E_{i-1}^e(n)-|\Delta_i(n)|^2/E_{i-1}^r(n-1) \tag{4m}$$

$$E_i^r(n)=E_{i-1}^r(n-1)-|\Delta_i(n)|^2/E_{i-1}^r(n) \tag{4n}$$

$$\gamma_{i-1}(n-1)=\gamma_{i-2}(n-1)+|r_{i-1}(n-1)|^2/E_{i-1}^r(n-1). \tag{4o}$$

$$a_i^{(i)}(n)=K_i^r(n) \tag{4p}$$

$$b_0^{(i)}(n)=K_i^e(n) \tag{4q}$$

$$\left.\begin{aligned} a_k^{(i)}(n)&=a_k^{(i-1)}(n)+K_i^r(n)b_{k-1}^{(i-1)}(n-1) \\ b_k^{(i)}(n)&=b_{k-1}^{(i-1)}(n-1)+K_i^e(n)a_k^{(i-1)}(n) \end{aligned}\right\} \quad 1\leq k\leq i-1. \tag{4r, 4s}$$

Least Squares Lattice Variables[b]

Lattice Parameter[c]	Symbol
Number of time samples (iterations)	N
Filter order	p
Time variable	(n)
Stage variable, current stage	i
Stage variable, lower stages	k
Input time sample	$x(n)$
Gain	$\gamma_{i-2}(n-1)$
Fade factor	$(1-\alpha_{CLSL})$
Step size parameter	$\Delta_i(n)$
Forward predictor coefficient vector	$\begin{cases} a_k^{(i)}(n) \\ 1 \leq k \leq i-1 \end{cases}$
Backward predictor coefficient vector	$b_k^{(i)}(n)$
Highest (ith) forward predictor	$a_i^{(i)}(n)$
Previous vector of backward predictors	$b_{k-1}^{(i-1)}(n-1)$
Forward power	$E_i^e(n)$
Backward power	$E_i^r(n)$
Forward PARCOR coefficient	$K_i^e(n)$
Backward PARCOR coefficient	$K_i^r(n)$
Forward error	$e_i(n)$
Backward error	$r_i(n)$

[a] from Hodgkiss & Presley (1982, pp. 331-332)

[b] adapted from Hodgkiss & Presley (1982, pp. 331-332)

[c] Variables pertain to ith stage unless otherwise noted.

DISTRIBUTION LIST FOR TM 83-90

Commander (NSEA 0342)
Naval Sea Systems Command
Department of the Navy
Washington, DC 20362 — Copies 1 and 2

Commander (NSEA 9961)
Naval Sea Systems Command
Department of the Navy
Washington, DC 20362 — Copies 3 and 4

Defense Technical Information Center
5010 Duke Street
Cameron Station
Alexandria, VA 22314 — Copies 5 through 10